计算机工程
Computer Engineering
ISSN 1000-3428,CN 31-1289/TP

《计算机工程》网络首发论文

题目：基于 CNN 与有限状态自动机的手写体大写金额识别
作者：闫茹，孙永奇，朱卫国，李宇霞
DOI：10.19678/j.issn.1000-3428.59118
网络首发日期：2020-10-15
引用格式：闫茹，孙永奇，朱卫国，李宇霞．基于 CNN 与有限状态自动机的手写体大写金额识别．计算机工程. https://doi.org/10.19678/j.issn.1000-3428.59118

网络首发时间：2020-10-15 10:39:57
网络首发地址：https://kns.cnki.net/kcms/detail/31.1289.TP.20201015.0951.005.html

计 算 机 工 程
Computer Engineering

基于 CNN 与有限状态自动机的手写体大写金额识别

闫茹[1] 孙永奇*[1] 朱卫国[1] 李宇霞[1]

（1 北京交通大学计算机与信息技术学院 北京 100044）

摘　要：手写票据识别是模式识别中的研究难点之一，由于手写体风格多样、票据背景复杂，识别准确率很难达到令人满意的效果。大写金额作为票据中最重要的部分，其识别准确率是手写票据自动识别的关键。本文对基于分割的手写大写金额的识别及后处理方法进行研究。首先，在利用过分割和组合过分割项得到单个字符后使用卷积神经网络(Convolutional Neural Networks, CNN)对单字符进行识别。然后，通过对字符进行分类、定义各类字符之间的逻辑关系构造出用于语法检测的有限状态自动机。再利用语法自动机在识别结果中选择出符合语法规则的字符串，并在路径搜索中利用语法自动机优化搜索性能。最后，在后处理中利用自动机对模糊字符进行预测，以纠正卷积神经网络的识别错误。实验结果表明，结合语法自动机校验和模糊字符预测的大写金额文本行识别准确率达到了 96.6%。

关键词： 卷积神经网络；有限状态自动机；手写票据识别；大写金额；光学字符识别；模式识别

开放科学（资源服务）标识码（OSID）：

Chinese Handwritten Legal Amount Recognition Based on CNN and Finite State Automata

YAN Ru[1], SUN Yong-qi*[1], ZHU Wei-guo[1] and LI Yu-xia[1]

(1 School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044)

Abstract: Handwritten Bank Check Recognition has been a tough problem in pattern recognition. With the challenging of different handwritten styles, noises and the complex background of bank checks, the recognition performance of handwritten bank check is still not very satisfactory. Chinese legal Amounts is the most important part of bank check, and its recognition accuracy is the key to automatic processing of handwritten bank check images. In this paper, we focus on the methods of segmentation-based recognition and post-processing. Firstly, the characters which are obtained from the combination of over-segmented items are recognized by convolutional neural networks. Secondly, an automaton for grammar detection is constructed by classifying the characters and defining their logical connection. Thirdly, the automaton is applied to reject grammatically incorrect amounts of the recognition results and optimize the performance of paths search. Finally, the automaton is used to infer the characters that are difficult to recognize by the convolutional neural network in post-processing. The experimental results show that the accuracy of line recognition of the legal amount can achieve 96.6% with the verification and prediction of grammatical automaton.

Keywords: Convolutional Neural Network; Finite State Automata; Handwritten bank check recognition; Chinese Legal Amounts; Optical Character Recognition; Pattern Recognition

DOI: 10.19678/j.issn.1000-3428.59118

0 概述

票据是银行、单位和个人办理支付结算和现金收付业务的重要依据，是记载经济业务和明确经济责任的一种书面证明。票据的数量繁多且非常重要，操作人员稍有不慎便可能造成很大的经

基金项目：国家自然科学基金 NSFC(No.61572005，No.61672086，No.61272004).

作者简介：闫茹（1994-），女，硕士研究生，研究方向为图像处理与深度学习；

孙永奇（通信作者），男，教授；朱卫国，男，博士研究生；李宇霞，女，硕士研究生。**E-mail：**yqsun@bjtu.edu.cn

济损失，因此利用光学字符识别(Optical Character Recognition, OCR)技术对票据进行自动识别就显得非常重要。作为票据中最重要的部分，金额的识别尤为关键，不能出任何差错。目前，手写票据中大写金额部分的识别准确率达不到令人满意的效果，原因在于：(1) 用户在书写时风格多样，字符内部、字符之间的空隙大小不均，字与字之间可能粘连严重；(2) 手写票据有一大部分是复写数据，字迹不清楚、笔画缺失，噪声难以完全去除。此外，汉字文本行的识别涉及整条文本行图像的去噪、准确切分、单字识别以及路径最优化选择等多项复杂任务，每一部分的性能都会影响到最后的识别结果。因此，对手写票据中大写金额的准确识别仍然是一个亟待解决的问题。

为提高大写金额的识别准确率，需要对初步的识别结果做进一步的检查调整。由于票据中大写金额的填写具有严格的语法要求，所以很适合使用有限状态自动机对识别结果进行检验。而且，大写金额字符数量少、内容明确，很适合构造有限状态自动机。此外，使用有限状态自动机可以理清字符间的语法逻辑，快速而准确地拒绝有语法错误的字符串，并定位到错误字符位置。

本文提出了一种基于卷积神经网络和有限状态自动机的手写体大写金额识别模型。在利用卷积神经网络对单个字符进行识别后，使用有限状态自动机检验大写金额的语法，并在票据识别结果的多条路径中选择出更加准确的符合语法规则的字符串。在路径搜索过程中，结合语法自动机的前缀判断功能优化搜索性能，对不必要的搜索路径剪枝以提高搜索的速度和精度。此外，还利用语法自动机对模型识别不出来的模糊字符进行预测以提高识别的准确率。

1 相关研究

1.1 手写大写金额识别

近几十年来，很多研究者致力于中文票据的识别[1, 2]，有专注于票据中手写小写金额的识别[3]，也有关注大写金额的处理和识别[4]。Lou[5]等人提出了一种偏旁识别器，使用字符宽度模型来确定分割线的预测位置。Chi[6]等人提出了一种基于隐马尔可夫模型的中国法定金额识别方法。在训练阶段，从滑动窗口中提取梯度特征，用单字符图像训练字符 HMM。在识别阶段，利用句子 HMM 对文本行图像进行分割，句子 HMM 由字符 HMM 根据严格的语言模型构造。Liu[7] 使用笔画序列信息和中文字符的八方向特征进行在线手写汉字识别。Gan[8]等人将文字表示为几何图形，同时保留空间结构和时间顺序，使用空间图卷积网络对字符图进行分类。Wang[9]等人使用深度学习自动编码器提取孤立大写金额汉字字符的特征，以提高识别字符准确率。Yu[10]等人提出了一种大写金额语法检查器，通过列举不符合正确语法的静态规则并结合动态判断来拒绝错误的识别结果。尽管对手写大写金额识别的研究取得很大的进展，但识别准确率仍然有可提高的空间。传统的基于过分割、组合、识别、路径选择、校验的识别方法已经被证实可以很好的应用于手写字符串的识别[11-13]。近期，有些学者提出了不需要分割字符的端到端的序列识别方法[14,15]，但这些方法大多对于规范的印刷体文本具有较好的性能，但对于复杂背景下的手写字符串识别还不能达到很好的效果。

本文将采用基于分割的框架对手写票据中的大写金额进行识别。利用过分割和组合过分割项来得到单个字符，并然后使用卷积神经网络(Convolutional Neural Networks, CNN)对单字符进行识别。然后，将构建语法自动机用于在路径搜索中选择出符合语法规则的识别结果，并在后处理中用于对模糊字符进行预测。

1.2 有限状态自动机

上世纪五十年代，美国语言学家乔姆斯基等人建立了形式文法和自动机之间的联系[16]，使得自动机可以有效地表达各种语法规则。由于确定

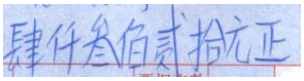
性的自动机在算法执行速度上与输入字符串的长度呈线性关系，且空间复杂度低，所以被广泛应用于自然语言处理领域中的语法规则建模[17,18]。Rachman[19]使用有限状态自动机将单词切分为音节，根据音节将文本转换为声音。有限状态自动机是具有离散输入、输出系统的一种数学模型，它定义了有限个内部状态和状态之间的转移函数。自动机按序接收输入信号，该信号在内部状态之间发生转移。对于不符合自动机转移条件的输入将被拒绝；符合规则的输入最终会转移至输出状态，表明该输入通过了检验。在实际的应用中，自动机的状态可能数以万计，而且自动机的规模也是逐渐变大的。因此，就会存在多余的状态或多余的转换弧，这些多余的状态和转换弧不仅降低计算速度，还浪费存储空间。这时，就需要对有限状态自动机进行最小化，使其没有多余的状态，并且没有两个状态是相互等价的[20]。在第 3 节中，将通过对字符进行合理分类来保证所设计的有限状态自动机没有冗余状态和重复的状态转换，以保证其最小化。

2 票据识别

票据中手写大写金额的识别主要包括：票据预处理及文本定位、字符图像的过分割、组合过分割项并识别以及路径搜索和校验等。

2.1 票据预处理及文本定位

手写票据图像是通过扫描仪获得的 RGB 三通道图像。为了方便对图像进行处理，减少计算量，需要对图像进行灰度化处理。由于票据的尺寸、文本布局和扫描方式的不一致性，所以会造成大写金额的位置在票据中的位置有所差异。因此，首先使用能在复杂背景下对水平文本进行检测的 CTPN 算法[21]进行大写金额部分的定位。文本定位之后，通过如下步骤进行去噪处理：（1）使用 Mask 匀光法[22]结合红色（R）通道去除的方法将图像中的红线去除，再转成灰度图像；（2）使用自适应阈值方法（OTSU）获得阈值，然后将大于该阈值的像素灰度值设置为 255，其余像素灰度值保持不变；（3）采用联通区分析对上下位置的小联通区进行去噪处理。大写金额定位后预处理前后的图像如图 1 所示。



(a) 大写金额定位结果

(a) The locating result of the legal amount

(b) 去噪后的图像

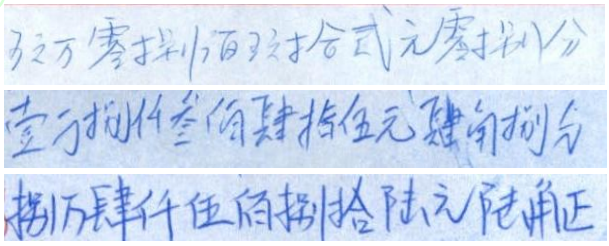
(b) The denoised image

图 1 票据预处理

Fig.1 Processing of a bank check

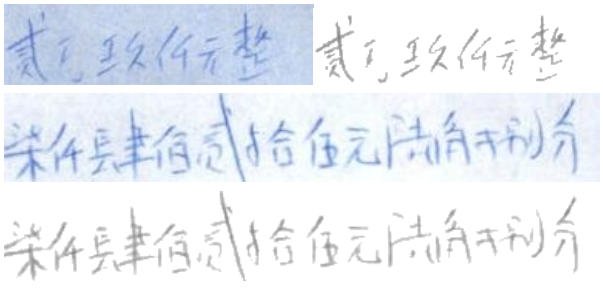
2.2 过分割与组合

手写汉字书写随意、风格多样，文本行字符之间粘连严重如图 2(a) 所示。因此，需要先采用过分割的方式进行文本行的切分。过分割算法将文本行图像尽可能分开，分割得到的每一个图像可能是单个字符或者是字符的一部分，这些图像被称为过分割项。



(a) 粘连文本行

(a) Text lines images with connected characters



(b) 笔画缺失的文本行

(b) Text lines images with strokes lost

图 2 票据样本

Fig. 2 Samples of bank checks

本文采用的分割算法主要分为以下三个步骤：(1) 通过联通区分析，将文本图像按照笔画或者部件进行分割；(2) 根据联通区域的重叠度进行初合并，调整合并结果，去除重复区域[23]；(3) 最后，检测粘连笔画并且对粘连项进行分割，通过计算过分割项的高度以及文本行高设定合适的宽高比阈值，高于阈值的项被认为可能存在粘连。对于潜在的粘连，沿着水平方向遍历，通过统计垂直方向上连续像素点的个数，再结合笔画宽度阈值和位置信息找到笔画为“-”的区域进行过分割处理。

单个字符的识别之前还需要对上述过分割项进行组合。使用的组合方案为：对过分割项的合并采用遍历方式，从第一个过分割项开始，依次与其后的过分割项进行合并，并记录好开始和结束位置。汉字“捌”由 4 个部件组成，是大写金额汉字中部件数目最多的文字。因此定义单个字符的最大合并过分割项 maxC = 4。

由于有些字符不需要合并 4 个过分割项，因此合并需要对过分割项进行类型判断，动态调整合并项的数目。本文根据过分割项的宽度 h 和高度 w 以及过分割项的最大宽度 maxW 和图像的行高 lineH 将过其划分为三类：

第一类 C_1：若过分割项的 w < maxW/th_1 且 h < lineH/th_2 则定义为小型。该类可以是单个瘦长型字符，也可以是字符中的单个部件，或者是部件的附属部分。

第二类 C_2：若过分割项的 $h/w>th_3$ 或者 w/lineH>th_4 则定义为中型，该类可以是单个字符，也可能是单个字符中的主要部件。

第三类 C_3：其余情况定义为大型，该类为单个字符。

其中，阈值 th_1、th_2、th_3 和 th_4 通过大量的优化实验来确定，票据识别中阈值的取值分别为 th_1=3, th_2=4, th_3=1.2, th_4=0.6。在合并的过程中当组合了一个 C_3 后，不再连续组合 C_3 或 C_2；每个过分割项最多组合 3 个 C_2; C_1 总是与其它项组合，如果合并项宽度大于设定阈值，则不组合。根据上述规则便可得到待识别的合并项。

2.3 字符识别

基于深度学习直接从原始数据中学习判别式特征来进行字符分类的方法已经取得了很大的进展。因此本文使用深度卷积神经网络(CNN)来训练字符分类器。

表 1 CNN 网络结构及其参数

Tab.1 Structure of CNN network and its parameters

Items	Size	Parameters（s, p）
input	1×64×64	
Convolution	64×3×3	(1, 1)
MaxPooling	64×2×2	(2, 0)
Convolution	128×3×3	(1, 1)
MaxPooling	128×2×2	(2, 0)
Convolution	256×3×3	(1, 1)
MaxPooling	256×2×2	(2, 0)
Convolution	512×3×3	(1, 1)
MaxPooling	512×2×2	(2, 0)
Convolution	512×3×3	(1, 1)
MaxPooling	512×2×2	(2, 0)
FullyConnected +Dropout	4096	dropout=0.5
FullyConnected +Dropout	4096	dropout=0.5
FullyConnected +Softmax	21	

本文采用 8 层的卷积网络，包含 5 个卷积层和 3 个全连接层。卷积层均采用 3×3 的卷积核，卷积步长为 1，padding 为 1。所有卷积层后接一个空间最大池化层 MaxPooling 来降低特征维度，同时保证图像的平移不变性，以增强分类的鲁棒性。池化窗口为 2×2，步长为 2。为了加快收敛速度和增强分类效果，在卷积层后均加入 BatchNorm 层。在全连接层后接 Dropout 层来增强模型的泛化性能。使用 RELUs 作为除最后一个全连接层外的所有卷积层和全连接层的激活函

数。最后一层使用 SoftMax 来进行分类。网络的输入为单通道的 64×64 的灰度图。为了便于计算，设置图像背景的灰度级为 0，前景为[1, 255]。CNN 的配置信息如表 1 所示。

3 构造大写金额语法自动机

使用有限状态自动机检验大写金额语法时将主要解决以下两个关键问题：(1) 根据大写金额语法规则对字符进行分类，通过定义合理的状态，构造有限状态自动机；(2) 根据构造的有限状态自动机，设计大写金额字符串的匹配算法。本节将对这两个问题进行详细描述。此外，将在 3.4 节和 3.5 节介绍所提出的语法自动机在前缀判断和模糊字符预测方面的应用。

3.1 大写金额语法规则

大写金额的主要语法规则如下[24]：

（1）中文大写金额的数字字符应该使用规范的字符，如壹、贰、叁、肆、伍、陆、柒、捌、玖、拾、佰、仟、万、亿、元（圆）、角、分、零、整（正）等字样。

（2）中文大写金额数字到“元”为止的，在“元”之后写“整”(或“正”)字；到“角”为止的，在“角”之后可以写也可以不写“整”(或“正”)字；大写金额数字有“分”的，“分”后面不写“整”(或“正”)字。

（3）如果大写金额对应的阿拉伯数字金额中有“0”时，大写金额应该符合汉语语言规律和金额数字构成：

①阿拉伯数字中间有“0”时，中文大写金额要写“零”字。如￥1409.50，应写成人民币壹仟肆佰零玖元伍角（整）。

②阿拉伯数字中间有连续多个“0”时，中文大写金额中间可以只写一个“零”字。如￥6007.14，应写成人民币陆仟零柒元壹角肆分。

③阿拉伯数字万位（元位）是“0”，或者数字中间有连续多个“0”且万位（元位）也是“0”，但千位（角位）不是“0”时，中文大写金额中可以只写一个“零”字，也可以不写。如￥107001.53，应写成人民币壹拾万零柒仟零壹元伍角叁分，或者写成人民币壹拾万柒仟零壹元伍角叁分；又如￥1680.32，应写成人民币壹仟陆佰捌拾元零叁角贰分，或者写成人民币壹仟陆佰捌拾元叁角贰分。

④阿拉伯金额数字角位是“0”，而分位不是“0”时，对应的大写金额“元”字后面应写“零”字。如￥16409.02，应写成人民币壹万陆仟肆佰零玖元零贰分。

3.2 构造有限状态自动机

由于大写金额字符组合的多样性，如果按单个字符作为有限状态自动机的状态，自动机将变得很大，而且由于多个状态具有相同的状态转移关系，状态之间的转换会变得冗余且复杂。因此，通过对字符进行分类可以得到最小化的有限状态自动机，同时准确的分类也有助于简化字符组合的复杂度从而使得字符间的逻辑关系更加清晰。本文根据每个字符在金额中的意义将输入字符集合划分为以下五类（括号内为集合名字）：

（1）数字类(Number)：壹，贰，叁，肆，伍，陆，柒，捌，玖；

（2）数字单位类(Unit)：拾，佰，仟，万；

（3）金额单位类(Amount)：元，圆，角，分；

（4）零类(Zero)：零；

（5）结束字符类(Only)：正，整。

自动机的每个状态以类别名称的首字母作为状态名称来表示自动机所处的状态，则上述集合对应的自动机状态分别为 *N*，*U*，*A*，*Z*，*O*。通过分析大写金额的结构可以看出，在“元（圆）”之前的金额表示是由数字字符和数字单位字符组成的，而在“元（圆）”之后则不会用到数字单位字符。因此，为了更加清晰地表达自动机状态之间的逻辑关系，将 Number 类和 Zero 类进一步划分为以下几类：

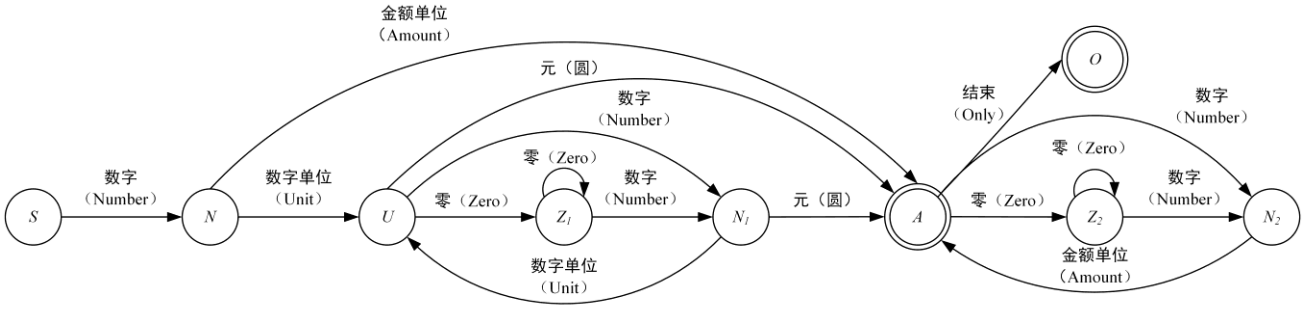


图 3　状态转换图

Fig.3　States transition diagram

（1）*N* 类：由于首字符缺乏上下文信息，因此将第一个数字字符归为该类。例如对于字符串“伍角叁分”或者“伍拾元”，当自动机读取第一个字符“伍”时，由于没有足够的信息将其分类，所以自动机进入 N 状态。

（2）N_1(N_2)类：在“元（圆）”之前（后）且不是首字符的数字类字符。

（3）Z_1(Z_2)类：在“元”之前（后）的零类字符。

用于大写金额语法检测的有限状态自动机 *M* 定义为一个有序五元组：

$M = \{ Q, \sum, \delta, q_0, F \}$，

其中，

$Q = \{ S, N, U, Z_1, Z_2, N_1, A, N_2, O\}$是自动机所有状态的集合；

$\sum$ = {壹，贰，叁，肆，伍，陆，柒，捌，玖，拾，佰，仟，万，亿，元，圆，角，分，零，整，正}为所有输入字符的集合；

δ 为状态转移函数: $Q \times \sum \rightarrow Q$；

$q_0 = S$ 是自动机的初始状态, $q_0 \in Q$；

$F = \{A , O\}$是所有终止状态的集合，$F \in Q$。

有限状态机的状态流程图为图 3 所示的有向图，可以看出它具有有限个状态节点，并且每个节点代表一个状态，每条有向边代表状态的转移方向。有向边上方标注的集合名称是状态转移时需要的字符类型，即外界对自动机的输入，自动机通过读取相应集合元素，推动状态的转移。

对于被检测的字符串，有限状态自动机依次读取其中的字符，依据匹配算法进行状态转移。每次读入一个字符发生一次状态转移，转移方向由当前所处状态和读入的字符共同决定。例如，对于 $U = \delta(N,$ Unit)，在自动机处在 *N* 状态时，当读入 Unit 字符集合中的一个单位字符时，自动机由 *N* 状态转移至 *U* 状态，发生一次状态转移。当自动机到达某个状态时表示该自动机接受了在此之前读入的所有字符组合成的字符串，而没有被接受的字符串表示不符合语法规则。因此，如果一个大写金额字符串可以从状态机的开始状态经过若干个中间状态后到达最终状态，则该字符串有效，否则无效。由于大写金额的规则规定可以以“角”，“分”和“整（正）”结尾，所以当字符串进入 *A* 状态时需要判断是否结束。因此，状态 *A* 既可以作为结束状态，也可以在还有字符读入时，根据转移函数转移至下一状态。

以上对有限状态自动机进行了简要描述，但在实际应用过程中情况要复杂很多。数字单位和金额单位中的字符具有紧密的上下文关系，例如“叁拾”符合语法规则，但“叁拾伍佰元整”则不符合语法规则。这种类型的规则需要结合上文信息才能表示，本文将其称为动态规则。通过在自动机的状态内部添加动态约束，实现对大写金额语法规则的合理判断。

3.3 自动机动态约束

在第 3.2 节中，为了使自动机最小化，在设计状态时合并了自动机中的等价状态。其中，数

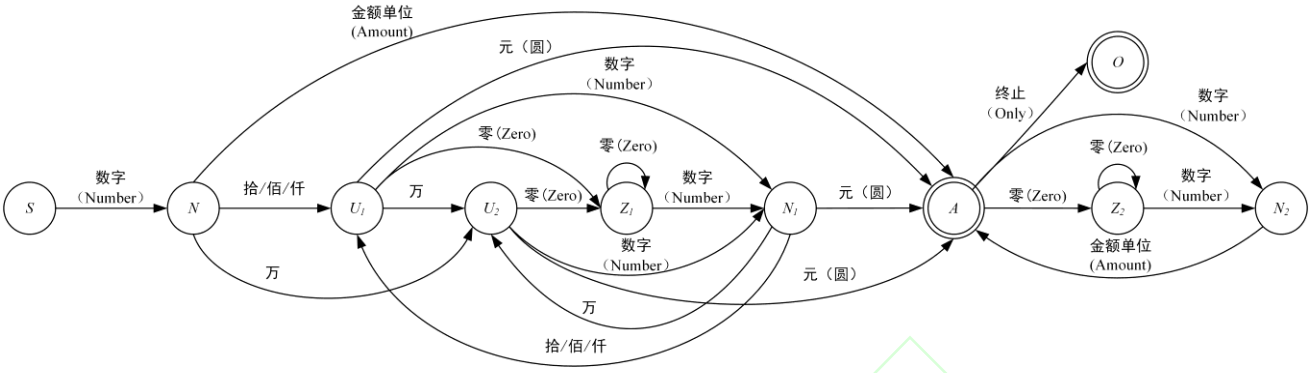


图 4　改进的状态转换图

Fig.4　The improved states transition diagram

字状态的 9 个数字字符是完全等价的，逻辑关系可以统一表示，所以状态内部不需要做约束。对于数字单位 Unit 和金额单位 Account 集合中的字符，需要结合语义信息在状态内部添加动态规则进行约束。本文根据单位状态发生转换时读入的字符将约束分为 4 类：数字类一般约束；数字类特殊约束；零类一般约束；零类特殊约束。为了方便对动态规则进行建模，定义以下符号：

（1）unit_all = {万，仟，佰，拾，元，角，分}，包含了数字单位 Unit 和金额单位 Account 集合中的元素。U_{index} 表示 unit_all 中各个字符对应的状态，$0 \le index \le 6$。例如，U_0 表示集合中的“万”字对应的自动机的状态，U_3 表示集合中的“拾”字对应的自动机的状态。

（2）wan_used 表示在自动机读过的字符串中“万”字出现过。

（3）unit_wan_used 表示数字单位集合 Unit 中的任意一个字符与“万”字符的组合是否出现过，例如“拾万”或“佰万”等。对于“元”字符，同样定义 unit_yuan_used 来表示类似的信息。

（4）m 和 n 分别表示当前单位状态下标以及下一单位状态的下标，且 $0 \le m < 6, 0 < n \le 6$。

具体的约束如下：

约束 3.1 数字类一般约束

if $m \neq 3$ then

$$U_n = f(U_m, number)$$

$$n - m = 1$$

endif

该约束是针对数字构造的一般规则，其中 f 为动态约束函数。用 index 表示 unit_all 中的状态，控制单位的移动。当且仅当读取一个数字字符时，数字单位向后移动一位，即 index++。所以该约束条件主要针对没有发生单位向前跃变的情形。当自动机读取字符串中数字字符的下一位单位字符时，如果单位字符与 index 指向的单位相同，则判定为正确的语法。否则，自动机返回 false 标志，表示该字符串被自动机拒绝，结束判断。但当 m = 3 时，对于合法的字符串“伍仟肆佰叁拾壹万”，在 U_m =“拾”且读入 Number 集合中的“壹”字符时，数字单位状态 U 中 index 由 3 变为 0（U_n =“万”），发生了单位的跃变，这时就需要对其做特殊约束。

约束 3.2 数字类特殊约束

if $m = 3$ then

　if *wan_used* = false then

$$U_0 \text{ or } U_n = f(U_m, number)$$

$$n - m = 1$$

　else

$U_n = f(U_m, number)$

$n - m = 1$

endif

endif

当 m = 3 时（U_m =“拾”），由于“万”字符的影响可能会发生单位的向前跃变。因此，此约束条件取决于“万”字符是否出现过。当“万”字符出现过时，表明万字段的数字已经处理完成，下一步按约束条件 3.1（没有发生跃变）处理即可。例如，当自动机已经读到字符串“壹万壹仟壹佰壹拾”的“拾”时，m = 3。再读入 Number 集合中的数字时，状态只能转移至 n = 4，即下一个单位字符只能为“元”。

当“万”字符没有出现过时，可能是该字符串的金额没有超过一万，也可能前面的数字单位是万字段的字符。对于前者，例如“壹仟壹佰壹拾壹元”，则不发生跃变；对于后者，如“壹仟壹佰壹拾壹万”，则发生跃变。这些都符合语法规则，需要根据接下来读入的数字单位做相应的选择。

约束 3.3 零类一般约束

if $m \neq 0$ && $m \neq 4$ then

$U_n = f(U_m, zero)$

$n - m \geq 1$

endif

该约束条件主要针对 3.1 节中第 3 条规定的第 2 款。当阿拉伯数字中间有连续多个“0”时，大写金额可以只写一个“零”。因此，结合数字构造规则在读入字符“零”时，index 至少需要向右移动一个单位，也可以移动多个单位，即 n–m ≥1。在读入下一个单位字符时，如果该单位满足此条件，则判定为语法正确，index 根据读入的字符进行相应赋值。否则，自动机返回 false，结束判断。当大写金额字符串为多个零时，index 按照零的个数进行移位。

约束 3.4 零类特殊约束

if $m = 0$ || $m = 4$ then

if (m = 0 && *unit_wan_used* = true) ||

(m = 4 && *unit_yuan_used* = true) then

$U_n = f(U_m, zero)$

$n - m \geq 0$

else

$U_n = f(U_m, zero)$

$n - m \geq 1$

endif

endif

该约束条件主要针对 3.1 节中第 3 条规定的第 3 款。该约束条件主要取决于数字单位与“万”或者“元”字符是否组合出现过。例如，对于“万”位，当 unit_wan_used = true 并且千位不为“零”时，该组合单位后面可以加零也可以不加零。在自动机中当读入字符“零”时，单位可以向后移动一位，或者不移动。当读入下一个字符时，满足此条件则判定为语法正确，否则自动机返回 false，结束判断。对于“元”位的处理也一样。

3.4 自动机在路径搜索中的应用

由于处理的数据是复写在票据上的大写金额字体，所以去噪之后这些字体更加不清晰，如图 2(*b*)所示，从而导致一些字符可能被分成三个以上的分割项。每个过分割项在尝试组合之后都存在一种或多种组合情况，每种组合情况还会有一个或多个识别结果。所有识别项都可以用四元组{*C*，*P*，*S*，*E*}表示，其中 *C* 为字符分类结果，*P* 为置信度，*S* 和 *E* 分别为合并项的开始和结束位置。本节将对首尾相连的识别项进行路径搜索，并将设计的语法自动机应用到动态搜索过程中，以提高路径搜索的精度和速度。

在路径搜索的过程中，把当前路径中的字符构成的字符串称为前缀字符串。前面所设计的自动机校验算法是针对完整的字符串，不能到达终止状态时该字符串被认为不符合语法规则。由于前缀字符串不是完整的字符串，所以需要对自动机进行适当调整以满足语法检验的需求。首先，

调整自动机的每个状态，使其可以接受标志字符 T，并可以由该状态直接进入终止状态。然后，在路径搜索过程中的前缀字符串后面添加判断标志 T。具体搜索步骤如下：

（1）对于每个合并项的识别结果按置信度 P 由高到低排序，当最大置信度高于阈值 T 时，保留前 n 项作为候选集。否则，认为该合并项的识别结果均不可靠，以模糊字符“卍”代替。接着，从 $S = 0$ 的合并项开始集束搜索，设置搜索宽度 $sw = 3$。

（2）使用自动机的前缀语法检测功能对搜索到的路径进行判断。当前缀字符串符合语法规则时，记录结束位置 E，并继续对开始位置为 $S = E + 1$ 的合并项进行搜索。否则，增加搜索宽度，即 $sw = sw + 1$。当所有的以 S 为开始的合并项都不符合语法规则时，将该位置的识别结果以模糊字符“卍”替代（C =“卍”），继续进行搜索。

（3）当结束位置 E 为最后一个合并项的结束位置时，就完成了路径搜索。由于整个搜索过程按照置信度优先搜索，因此以第一条符合语法规则的路径作为最后的识别结果。

当搜索路径中的模糊字符“卍”的个数大于 3 时，或者出现两个连续的模糊字符，就认为组合或识别结果不可靠，该搜索路径被剪枝。对模糊字符在规定范围内的字符串，暂不使用语法自动机对其进行检测。在模糊字符被预测出之后归为候选路径，模糊字符预测的详细过程在 3.5 节中进行介绍。

3.5 自动机在模糊字符预测中的应用

对于一些笔画缺失或者噪声去除效果不好的字符图像，卷积神经网络很难给出正确的识别结果。这类字符在路径搜索中以模糊字符“卍”替代。由于大写金额无论是数字单位还是金额单位都有规则的序位关系，因此可以利用自动机对这类模糊字符进行预测。

自动机每一次的字符读取都有确定的转移状态，当字符串中有单个模糊字符时，模糊字符必然出现在前一状态的下一个转移状态中。这时，就可以将下一状态的字符集作为该模糊字符的候选集。以字符串“壹仟零叁卍陆仟叁卍捌拾元整”为例，当自动机读到第一个“卍”字时，自动机处于 N_1 状态，从自动机的状态转换图可以看出该状态的下一状态只能是 M 和 U 两种情况。当下一状态为 M 时，“元”作为该模糊字符的预测候选项。当下一状态为 U 时，由于“仟”和 “零” 字符的出现，根据 3.3 节中的约束 3.1 可知，此时 index = 3，预测候选项为“拾”。同时，由于 wan_used = false，根据 3.3 节的约束 3.2 可知，“万”也是预测候选项。因此，第一个“卍”字符有 3 个候选项{“元”，“拾”，“万”}，形成 3 条候选路径。自动机继续读取到字符“仟”时，语法自动机就会把包含预测值“元”和“拾”的路径剪枝。最终得到唯一的候选“万”，完成了一次准确的预测。同理，第二个模糊字符将被准确预测为“佰”。最终，语法自动机将返回所有模糊字符的预测结果。

4 实验结果

为了验证以上算法的有效性，本节将进行两部分的实验：第一部分为验证自动机语法检测和模糊字符预测两个算法的实验，使用的数据为大写金额字符串；第二部分为这两个算法在手写票据中大写金额部分识别的应用实验，使用的数据为手写票据。

4.1 实验数据及评价指标

本文使用从中科院手写汉字数据集 CASIA-HWDB1.0-1.2[25]中挑出的 21 类字符集进行卷积神经网络（CNN）的训练。其中，训练集每类包含 800 张不同作者手写的单字符图片。实验采用的测试数据是人工标注的 2600 张手写票据。

实验中针对不同的应用场景，将采用不同的评价指标。在单独的模糊字符预测实验中，使用单字符预测准确率(Character Prediction Accuracy,

CPA)和整条文本的预测准确率(String Prediction Accuracy, SPA)作为评价指标。在识别票据中大写金额图像文本时，采用单字符识别准确率(Character Recognition Accuracy, CRA)和整条文本的识别准确率(Line Recognition Accuracy, LRA)来作为评价指标。CRA 按照识别结果和标注值的编辑距离进行计算。

4.2 自动机语法检测和模糊字符预测结果

为了评估有限状态自动机的语法检测效果，使用 2000 条大写金额字符串做实验，其中 1000 条具有语法错误。实验结果表明，语法检测自动机对正确字符串的接受率和错误字符串的拒绝率均为 100%。因此，语法自动机可以对识别结果进行非常可靠的语法检测。

在对模糊字符预测效果的实验中，使用 2600 条手写票据中大写金额部分的 label 作为测试数据来评估预测效果。其中，1300 条文本随机使用模糊字符“卍”替代标注字符串中的 1 个位置的字符。剩下的 1300 条，随机使用两个模糊字符“卍”替代标注字符串中的 2 个字符。替换位置为非数字位，由于数字位的 9 个字符不具备上下文语义，因此本节仅评估非数字位的预测效果。自动机对模糊字符的预测效果如表 2 所示：

表 2 模糊字符预测结果

Table 2 Results of prediction for unclear characters

实验类型	CPA（%）	SPA（%）
1 个模糊字符	99.5	99.5
2 个模糊字符	96.0	95.6

从实验结果可以看出，当字符串中仅有 1 个模糊字符时，自动机的预测准确率可以达到 99.5%。当字符串的中有两个位置的模糊字符时，自动机的预测准确率会下降到 96.0%左右。由于大写金额语法的灵活性，同一模糊字符处可能存在多种符合要求的预测结果，所以自动机的字符预测率很难达到 100%。但作为识别的后处理步骤，基于自动机的模糊字符预测功能可以弥补部分模型识别错误的情况，进而提高票据中大写金额的识别准确率。

4.3 票据识别结果

本实验主要对 2600 张手写票据的大写部分进行识别。为进一步验证自动机模糊字符预测的重要性，本节将进行消融实验。实验 1 为识别后没有模糊预测的结果，实验 2 为识别后再进行模糊字符预测的结果。实验结果对比如表 3 所示：

表 3 模糊字符预测对票据识别的影响

Table 3 The effect of prediction on bank checks

实验类型	CRA（%）	LRA（%）
1	95.3	91.4
2	98.2	96.6

从实验结果可以看出，将语法自动机应用于票据中大写金额的识别时，字符识别准确率可达 95.3%，文本行的识别准确率达到了 91.4% 。该结果也验证了在路径搜索中结合前缀语法检测自动机可以提高搜索的精度和速度。在加入模糊字符预测的处理之后，单字符的识别准确率达到了 98.2%，提升了 2.9%；而整条文本的识别准确率达到了 96.6%，提升效果更为显著，达到了 5.2%。实验结果表明，在大写金额的后续处理步骤中使用语法自动机对模糊字符进行预测，可以有效提升识别的准确率。

5 结束语

本文采用基于分割的框架对手写票据中的大写金额进行识别。首先，通过对字符进行分类、定义类与类之间的逻辑关系，构造出可以用于语法检测的有限状态自动机。然后，利用语法自动机在识别结果中选择出符合语法规则的字符串。在路径搜索阶段采用束搜索的策略，通过前缀语法检测对搜索路径进行剪枝，避免了路径的指数级增长，提高了搜索速度。针对复写票据不清晰、笔画缺失导致卷积神经网络字符识别不正确或者置信度很低的问题，提出了模糊字符预测算法。实验结果表明，结合语法自动机校验和模糊

字符预测的单字符的识别准确率达到 98.2%，整条文本的识别准确率达到了 96.6%。

但是，自动机的模糊字符预测仅针对具有上下文关系的非数字位的预测，对于无上下文关系的数字位还无法预测。在下一步研究工作中，将先使用语法自动机对其定位，然后根据偏旁部首或部件信息进行纠正。此外，还可以对语法自动机进一步改进，增加对具有错误语法的字符串进行纠正的功能。

参考文献

[1] HUANG Liangli, LI Shutao, and LI Liming. "Extraction of filled-in items from Chinese bank check using support vector machines." *International Symposium on Neural Networks*. Springer, Berlin, Heidelberg, 2007.

[2] WANG Song, MA Feng, XIA Shaowei. A Chinese bank check recognition system based on the fault tolerant technique[C]. Document Analysis and Recognition, 1997, 2: 1038-1042.

[3] LIU Dong, and CHEN Youbin. "A prototype system of courtesy amount recognition for Chinese Bank checks." *2012 10th IAPR International Workshop on Document Analysis Systems*. IEEE, 2012.

[4] TANG Hanshen, AUGUSTIN E, SUEN C Y, *et al*. Recognition of unconstrained legal amounts handwritten on Chinese bank checks[C]. 17th International Conference on Pattern Recognition: IEEE, 2004, 2: 610-613.

[5] LOU Zhen, YANG Jingyu, and JIN Zhong. "Recognition and checkout of legal amounts on chinese bank cheques." *2008 International Conference on Wavelet Analysis and Pattern Recognition*. Vol. 1. IEEE, 2008.

[6] CHI Bingyu，CHEN Youbin. Chinese Handwritten Legal Amount Recognition with HMM-Based Approach[C]. International Conference on Document Analysis and Recognition: IEEE, 2013:778-782.

[7] LIU Xin, HU Baotian, Chen Qingcai, et al. Stroke Sequence-Dependent Deep Convolutional Neural Network for Online Handwritten Chinese Character Recognition[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[8] Gan Ji, Wang Weiqiang, Lu Ke. Characters as Graphs: Recognizing Online Handwritten Chinese Characters via Spatial Graph Convolutional Network[J]. arXiv preprint arXiv:2004.09412, 2020.

[9] WANG Meng, CHEN Youbin, and WANG Xingjun. "Recognition of handwritten characters in chinese legal amounts by stacked autoencoders." *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014.

[10] YU M L, KWOK P C K, LEUNG C H, et al. Segmentation and recognition of Chinese bank check amounts[J]. International Journal on Document Analysis and Recognition, 2001, 3(4):207-217.

[11] ZHANG Xuyao, Bengio Y, and LIU Chenglin. "Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark." *Pattern Recognition* 61 (2017): 348-360.

[12] WANG Qiufeng, YIN Fei, and LIU Chenglin. "Integrating language model in handwritten Chinese text recognition." *2009 10th International Conference on Document Analysis and Recognition*. IEEE, 2009.

[13] WU Yichao, YIN Fei, and LIU Chenglin. "Evaluation of neural network language models in handwritten Chinese text recognition." *2015 13th International Conference on Document Analysis and Recognition*

(ICDAR). IEEE, 2015.

[14] SHI Baoguang, BAI Xiang, and YAO Cong. "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition." *IEEE transactions on pattern analysis and machine intelligence* 39.11 (2016): 2298-2304.

[15] MESSINA R, LOURADOUR J. Segmentation-free handwritten Chinese text recognition with LSTM-RNN[C]. International Conference on Document Analysis and Recognition: IEEE, 2015: 171-175.

[16] CHOMSKY N. Three models for the description of language[J]. IRE Transactions on Information Theory, 1956, 2(3):113-124.

[17] XIA Wuji, HUA Quecairang. Automatic translation between Arabic numerals and Tibetan numerals based on finite state automata[J]. 2018, 40(3):550-554. (in Chinese)

夏吾吉, 华却才让. 基于有限状态自动机阿拉伯数字与藏文数词自动翻译[J]. 计算机工程与科学, 2018, 40(3):550-554.

[18] GREGHI J G, MARTINS E. Semi-automatic generation of extended finite state machines from natural language standard documents[C]. IEEE International Conference on Dependable Systems and Networks Workshops: IEEE, 2015:45-50.

[19] Rachman F H, Solihin F. Finite State Automata Approach for Text to Speech Translation System in Indonesian-Madurese Language[C]//Journal of Physics: Conference Series. IOP Publishing, 2020, 1569(2): 022091.

[20] SUN Yuqiang, LI Yuping, WANG Haiyan, et al. Parallel processing of minimization algorithm for determination finite automata[J]. Computer Science,2008, 35(1)：298-300. (in Chinese)

孙玉强，李玉萍，王海燕，陈继光．确定有限自动机最小化算法的并行处理．计算机科学，2008，35(1)：298-300.

[21] TIAN Zhi, HUANG Wei Ling, HE Tong, *et al*. Detecting Text in natural image with connectionist text proposal network[C]. European Conference on Computer Vision: Springer, 2016:56-75.

[22] HAN Yutao. Research on Key Technology of Color Consistency Processing for Digital Ortho Map Mosaicing[D]. WuHan University,2014.

韩宇韬. 数字正射影像镶嵌中色彩一致性处理的若干问题研究[D]. 武汉大学, 2014.

[23] LIU Chenglin, YIN Fei, WANG Da Han, WANG Qiufeng. CASIA online and offline Chinese handwriting databases[C]. International Conference on Document Analysis and Recognition, 2011:37–41.

[24] Ministry of Finance Order No. 98 of the People's Republic of China. Accounting basic work specification [EB/OL]. (2019-03-14) http://tfs.mof.gov.cn/zhengwuxinxi/caizhengbuling/201903/t20190315_3193919.html. (in Chinese)

中华人民共和国财政部令第 98 号. 会计基础工作规范 [EB/OL]. (2019-03-14) http://tfs.mof.gov.cn/zhengwuxinxi/caizhengbuling/201903/t20190315_3193919.html.

[25] WANG Qiufeng, YIN Fei, LIU Chenglin. Handwritten Chinese text recognition by integrating multiple contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(8): 1469-1481.