基于 Viterbi 算法的粘连断裂印刷体数字行切分识别方法

殷绪成 [1,2] 刘昌平 [1]

摘 要 粘连断裂字符行的切分识别，是很多 OCR 实际应用中存在的主要困难之一. 本文针对粘连断裂的印刷体数字行，提出了一种基于 Viterbi 算法的切分识别方案，该方案采用两次切分识别的层次型结构. 在第二次切分识别过程中，首先，在候选切分点区域，结合灰度图像与二值轮廓信息，采用基于 Viterbi 算法搜索的非直线路径进行切分，得到有效的切分路径；然后，结合分类器输出的可信度，采用 Viterbi 算法来合并前面得到的候选切分图像块，进行动态切分与识别. 实际的金融票据识别系统实验表明，本文提出的印刷体数字行切分识别方法能够较好的克服字符行的粘连与断裂情况，提高了识别系统的识别率和鲁棒性.

关键词 字符切分, OCR, 粘连断裂字符, Viterbi 算法, 印刷体数字行
中图分类号 TP391.4

A Segmentation and Recognition System for Touching and Broken Numeral Strings Based on Viterbi Algorithms

YIN Xu-Cheng[1,2] LIU Chang-Ping[2]

Abstract Currently, in many OCR applications, it is difficult to segment and recognize touching and broken characters. In this paper, a segmentation and recognition system based on Viterbi algorithms is proposed to solve such a problem for touching and broken machine-printed numeral strings. This system includes two steps of segmentation and recognition. In the second step, first, a segmentation method is adopted to find the character nonlinear segmentation paths by combining gray scale and binary information based on a Viterbi algorithm; then, a recognition method of using a Viterbi algorithm is adopted to dynamically combine and recognize the character candidates with their reliabilities generated from the recognizer. Some experiments on a financial document analysis and recognition system indicate that this Viterbi algorithms based method is efficient for segmentation and recognition of touching and broken numeral strings, and enhances the accuracy and robustness of the recognition system.

Key words Character segmentation, OCR, touching and broken characters, the Viterbi algorithm, machine-printed numeral strings

1 引言

在实际应用中，OCR 的瓶颈不再是分类器的设计问题，而主要取决于字符切分，特别是粘连断裂字符行的切分问题. 对于干净的印刷体文本行，简单的基于字符特征的切分就能够达到实用效果；而对于有噪声的印刷体、限制型手写体、和无限制的手写体，则需要依次采用更加复杂的切分方法. 一般来说，对于复杂的切分问题，字符切分与识别是结合在一起考虑的；也就是说，现行很多实用的字符切分方法都是基于字符特征和分类器识别的混合型方法[1]. 一般的字符切分都是基于二值图像的；但是，在比较复杂的情况下，仅仅利用二值信息往往不能得到满意的切分效果. Lee 等人利用灰度图像来进行切分与识别[2]；而在文献 [3] 中，研究者提出了结合灰度图像和字符二值轮廓信息的字符切分识别方法.

对于印刷体数字行的切分，由于打印机的不同、打印字体的不同、以及打印油墨的浓淡不同，容易产生粘连字符和断裂字符；而且字符宽度和高度信息在实际应用中也是变化不定的. 未能准确分割粘连与断裂字符是产生识别错误的主要原因之一，这已经成为实际应用中的主要瓶颈，而一些经典切分识别方法[4] 很难较好的解决这些问题.

Viterbi 算法是一种简单有效的解码方法[5,6]，可以看作为图结构中的最短路径搜索动态规划方法. Viterbi 算法提供了一个统一框架，能够较好的描述路径搜索的动态特性； 而且可以方便的结合先验知

收稿日期 2005-4-15 收修改稿日期 2005-10-11
Received April 15, 2005; in revised form October 11, 2005
1. 中国科学院自动化研究所文字识别工程中心 北京 100080 2. 中国科学院研究生院 北京 100049
1. Character Recognition Engineering Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100080 2. Graduate School of Chinese Academy of Sciences, Beijing 100049
DOI: 10.1360/aas-007-0315

识与规则. 该算法已经成熟的应用于多种模式识别问题中，其中最常见的就是语音识别[7] 和文本识别[4,8].

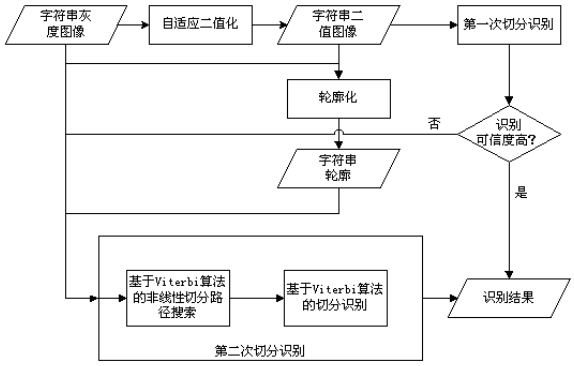

图 1 切分识别系统流程图
Fig. 1 The frame of our system

本文针对金融票据识别系统中存在大量粘连断裂字符的情况，提出了一种基于 Viterbi 算法的印刷体数字行切分识别方案. 在我们的识别系统中，采用层次型切分识别方法，系统流程如图 1 所示. 第一次切分识别直接在二值图像上进行，首先利用垂直差分投影分析[9]，确定初始候选切分点，然后进行识别；对识别可信度比较高的字符进行宽度估计，并利用这个宽度阈值对字符行再次进行切分识别. 如果第一次切分识别的可信度都比较高，则整个识别系统结束；否则进行第二次切分识别处理. 第二次切分识别结合灰度和二值信息进行，首先在灰度图像上，利用垂直差分投影信息，并结合第一次切分识别的结果，确定初始候选切分点；为了处理字符粘连情况，利用 Viterbi 算法[5]，结合二值及轮廓信息在灰度图像上搜索定位非直线的切分路径；最后，为了解决字符粘连断裂问题，再次利用 Viterbi 算法，在识别可信度的基础上动态合并这些非直线切分路径得到的候选字符块，得到最终的结果.

本文余下的内容安排如下. 第 2 节介绍了本识别系统所用的二值化方法和字符分类器设计；第 3 节分析了第一次切分识别过程；第 4 节重点介绍了基于 Viterbi 算法的第二次切分识别过程，最后是本文的总结和相关工作的展望.

2 二值化处理与分类器设计

2.1 自适应的二值化方法

Sankur 等人把二值化方法分为全局二值化方法和局部二值化方法[10]. 这些方法各有特点，全局二值化方法中性能表现好而且稳定的方法为 Otsu 方法[11]，而局部二值化方法中性能表现好而且稳定的方法则为 Niblack 方法[12]. 对于粘连断裂的灰度字符行图像，如果单纯的采用全局二值化方法，则进一步扩大了字符的粘连程度；如果采用单一的局部二值化方法，则会出现毛刺，不利于字符行的切分与识别. 在我们的系统中，采用自适应的二值化方法，结合 Otsu 方法和 Niblack 方法，得到的图像为两种二值化方法得到图像的 “与”. 设 $p(x,y)$ 为最后输出的二值化图像点 (x,y) 的值，$p_{Otsu}(x,y)$ 为 Otsu 方法得到的值，$p_{Niblack}(x,y)$ 为 Niblack 方法得到的值，则有

$$p(x,y) = p_{Otsu}(x,y) \quad \& \quad p_{Niblack}(x,y) \qquad (1)$$

其中，$p(x,y)=1$ 表示黑点 (前景字符)，$p(x,y)=0$ 表示白点 (背景). 当然，在图像二值化后，还要进行其它的一些预处理，如虑线、去噪和识别区域再定位等.

2.2 分类器设计

近邻法是一种简单有效的分类方法. 在我们系统中，采用基于聚类训练的近邻法.

在训练阶段，对训练样本进行诸如局部笔画方向和方向线素等方面的特征提取 (设共使用了 N_f 种特征)；然后聚类成形成多个子类. 在我们的系统中，共生成 100 多个子类. 显然，每一个类别有多个子类. 再利用主元分析 (PCA)，对这些特征向量进行降维处理. 最后把各子类的中心投影到 PCA 变换矩阵上，得到较低维数的子类表示模板.

在测试阶段，对于新来的样本，经过特征提取和降维，得到低维的特征向量；采用最近邻法进行分类，即比较这些向量与各子类模板，如果该样本离某个子类模板的欧式距离最小，则该样本属于该子类.

设共有 N 个子类模板，其序列为 $\{M_1, M_2, ..., M_N\}$，sx 为待识别的样本，$\{ftr_1(sx), ftr_2(sx), ..., ftr_{N_f}(sx)\}$，为降维后的特征向量序列，若

$$g_j(sx) = \min_i \|ftr_j(sx) - M_i\| \qquad (1 \le j \le N_f, \quad 1 \le i \le N) \qquad (2)$$

$$g_k(sx) = \min_j g_j(sx) \qquad (1 \le j \le N_f) \qquad (3)$$

则 sx 属于第 k 个子类.

对于分类器输出的可信度 $S(sx)$，可以由 $g_k(sx)$ 归一化至 [0,1] 得到

$$S(sx) = 1 - (g_k(sx) - g_{\min})/(g_{\max} - g_{\min}) \qquad (4)$$

其中 $g_{\max}$ 和 $g_{\min}$ 为 $g_k(sx)$ 可能出现的最大值和最小值，$S(sx)=1$ 表示最可信，$S(sx)=0$ 表示最不可信.

这样设计分类器的一个优点就是，对于测试的增量样本，如果其与各子类的距离大于一定的阈值，则可以把该样本相应的特征向量增加到子类模板中，形成新的子类模板. 这种增量学习的方式，在实际应用中是非常方便有效的.

3　第一次切分识别

第一次切分识别过程主要通过评估字符的宽度和识别可信度来进行. 如果第一次切分识别的可信度都比较高，则整个识别系统结束；否则进行第二次切分识别处理.

通过二值图像的垂直差分投影分析，投影值为极小的点为可能的切分点. 通过评估字符的大约宽度和各点之间的水平距离，进一步缩小点集的范围，设此时点集的水平坐标序列为 $\{x^1_{Mproj}, x^2_{Mproj}, ..., x^{N_{Mp}}_{Mproj}\}$，并进行垂直直线路径的切分，得到 $N_{Mp}-1$ 个候选字符 $\{C^1_{Mproj}, C^2_{Mproj}, ..., C^{N_{Mp}-1}_{Mproj}\}$.

切分后，利用 2.2 节的分类器进行识别，得到每一个候选字的识别可信度，得到的可信度序列为 $\{S^1_{Mproj}, S^2_{Mproj}, ..., S^{N_{Mp}-1}_{Mproj}\}$. 选择可信度大于某一阈值的候选字符，即

$$Set(C) = \{C^i_{Mproj}, S^i_{Mproj} > S^{THS}_{Mproj}\}, \quad (1 \le i \le N_{Mp}-1) \tag{5}$$

并计算 $Set(C)$ 中字符的平均宽度 W_{mean}. 然后，利用平均宽度 W_{mean}，并结合分类器输出的可信度，对印刷体数字行从左到右依次进行切分识别，从而得到相应的字符行识别结果，并得到第一次切分识别的切分点水平坐标序列 $\{x^1_{classify}, x^2_{classify}, ..., x^{N_c}_{classify}\}$.

4　基于 Viterbi 算法的切分与识别

4.1　基于 Viterbi 算法的非线性路径切分

对于粘连的字符行，在二值图像上进行的第一次切分识别，容易产生多切或少切的情况. 因此，利用图像的灰度信息是一种有益的补充方法. 文献 [2,3] 采用了基于灰度图像的动态规划非直线路径切分方法. 在我们的系统中，在灰度图像上，结合二值与字符轮廓信息，把象素点看作一种状态观测序列，利用 Viterbi 来计算搜索非直线切分路径. 该方法主要分为两步，首先是候选切分点的确认；然后在每一个候选切分点所在的切分区域内，进行基于 Viterbi 算法的非线性切分路径搜索.

4.1.1　候选切分点定位

通过灰度图像的垂直差分投影分析，投影值为极小的点为可能的切分点. 通过评估字符的大约宽度和各点之间的水平距离，进一步缩小点集的范围，设此时点集的水平坐标序列为 $\{x^1_{Gproj}, x^2_{Gproj}, ..., x^{N_{Gp}}_{Gproj}\}$. 结合第一次切分识别得到的切分点水平坐标序列 $\{x^1_{classify}, x^2_{classify}, ..., x^{N_c}_{classify}\}$，形成新的候选切分点水平坐标序列

$$Set(p) = \{x^1_{Gproj}, x^2_{Gproj}, ..., x^{N_{Gp}}_{Gproj}, x^1_{classify}, x^2_{classify}, ..., x^{N_c}_{classify}\} \tag{6}$$

这样可以完全的找出每一个可能切分点，而且这些切分点的数目在一定的范围内.

根据以下规则对这些候选切分点 $Set(p)$ 进行相关的合并和新增切分点操作：1) 如果两个切分点之间的距离小于某一阈值，则合并这两个点，合并得到的新切分点的水平坐标为这两个点坐标的平均值；2) 如果紧临的两点间的距离大于 $2 \times W_{mean}$ (W_{mean} 为第一次切分识别得到的平均字符宽度)，则在这两点之间再生成一个新的候选切分点，其水平坐标为这两个点坐标的平均值.

设最后得到的有效候选切分点为 N 个，其从左到右的水平坐标序列为 $\{x^1, x^2, ..., x^N\}$.

4.1.2　基于 Viterbi 算法的非线性切分路径搜索

对于上述 N 个候选切分点，每一个候选切分点有一个切分区. 设 x^n 为第 n 个候选点，则切分区 R^n 为

$$R^n = \{(x,y), \quad x^n - r_{ths} \le x \le x^n + r_{ths}, \quad top^n \le y \le bottom^n\} \tag{7}$$

其中，r_{ths} 近似等于字符笔画宽度，top^n 和 $bottom^n$ 分别为第 n 个候选点所在字符行区域的顶部和底部垂直坐标. 设上述切分区的高度为 H，则该切分区可用一个多阶 (H 阶) 的图来表示[2,3]，如图 2 所示. V_i 表示图中第 i 行所有点的集合，以 $< v_{i,j}, v_{i+1,k} >$ 表示第 i 行到 $i+1$ 行的边. 在该图中，一般以从第一行到最后一行以累计灰度最大 (255 表示白，0 表示黑) 的路径作为切分路径[2]；在文献 [3] 中，还利用了加权的累计灰度计算. 在图搜索过程中，如果图的边 $< v_{i,j}, v_{i+1,k} >$ 有效，则有 $j-1 \le k \le j+1$.

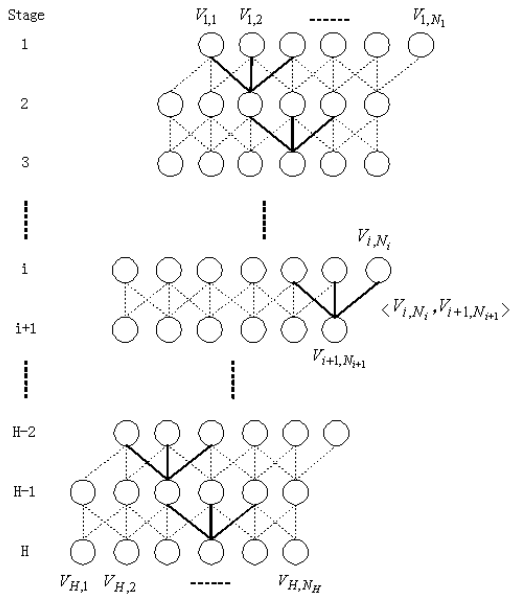

图 2 切分区的图表示[3]

Fig. 2 Graph representation in a segmentation region[3]

在实际应用中，我们发现，对于粘连情况严重的数字字符行，切分区中间区域通常粘连密度非常大，采用传统的自顶而下的搜索路径，往往不能很好的把粘连的字符切分开. 我们以 δ_n 表示切分区 R^n 的中间区域所对应的二值图像块的黑象素点密度

$$\delta_n = Num\{(x,y), x^n - r_{ths} \le x \le x^n + r_{ths}, top^n + H/4 \le y \le bottom^n - H/4\} \,/\, Num\{(x,y) \in R^n\} \tag{8}$$

其中 $Num()$ 表示该区域中黑点的个数.

另外，对于粘连字符，图的边 $< v_{i,j}, v_{i+1,k} >$ 仅限于 $j-1 \le k \le j+1$ 带来的效果也是有限的，在我们的系统中，采用 $j-2 \le k \le j+2$.

所以，本系统分为两条路径来进行切分：一条从切分区顶部开始，直达中部位置 mid^n_{top}；另一条从底部到中部位置 mid^n_{bottom}. 然后，再合并连接得到的两条路径，形成一条非直线的切分路径. mid^n_{top} 和 mid^n_{bottom} 的取值由 R^n 中间区域黑象素点的密度 δ_n 来决定

$$mid^n_{top} = \begin{cases} H/2 & \delta_n < \delta_{THS} \\ H/4 & \delta_n \ge \delta_{THS} \end{cases} \tag{9}$$

$$mid^n_{bottom} = \begin{cases} H/2 & \delta_n < \delta_{THS} \\ 3H/4 & \delta_n \ge \delta_{THS} \end{cases} \tag{10}$$

其中，δ_{THS} 为预先确定的中间区域黑象素点的密度阈值.

/* R^n上半部分 */

初始化： $i = top^n$;

$L(v_{top^n,k}) = 0, \ (top^n,k) \in R^n$;

$\Gamma(v_{top^n,k}) = \{\}, \ (top^n,k) \in R^n$;

递归： For $x^n - r_{THS} \le j \le x^n + r_{THS}$

For $j-2 \le k \le j+2$

$$l(v_{i,k}, v_{i+1,j}) = -[\log \frac{(1/5 + \alpha_{i,k} + \beta_{i,k})}{\sum_{m=j-2}^{j+2} (1/5 + \alpha_{i,k} + \beta_{i,k})} + \log \frac{p(i,k)}{255}];$$

$l(v_{i+1,j}) = \min_k l(v_{i,k}, v_{i+1,j})$;

$\tau(v_{i+1,j}) = \arg\min_k l(v_{i,k}, v_{i+1,j})$;

$L(v_{i+1,j}) = L(v_{i+1,\tau(v_{i+1,j})}) + l(v_{i+1,j})$;

$\Gamma(v_{i+1,j}) = \Gamma(v_{i,\tau(v_{i+1,j})}) \bigcup \{\tau(v_{i+1,j})\}$;

$i = i+1$;

终止： $L^*_{top-half} = \min_{x^n - r_{THS} \le j \le x^n + r_{THS}} L(v_{mid^n_{top},j})$;

$j^* = \arg\min_{x^n - r_{THS} \le j \le x^n + r_{THS}} L(v_{mid^n_{top},j})$;

路径求取： $\Gamma^*_{top-half} = \Gamma(v_{mid^n_{top},j^*})$.

/* R^n下半部分 */

初始化： $i = bottom^n$;

$L(v_{bottom^n,k}) = 0, \ (bottom^n,k) \in R^n$;

$\Gamma(v_{bottom^n,k}) = \{\}, \ (bottom^n,k) \in R^n$;

递归： For $x^n - r_{THS} \le j \le x^n + r_{THS}$

For $j-2 \le k \le j+2$

$$l(v_{i,k}, v_{i-1,j}) = -[\log \frac{(1/5 + \alpha_{i,k} + \beta_{i,k})}{\sum_{m=j-2}^{j+2} (1/5 + \alpha_{i,k} + \beta_{i,k})} + \log \frac{p(i,k)}{255}];$$

$l(v_{i-1,j}) = \min_k l(v_{i,k}, v_{i-1,j})$;

$\tau(v_{i-1,j}) = \arg\min_k l(v_{i,k}, v_{i-1,j})$;

$L(v_{i-1,j}) = L(v_{i-1,\tau(v_{i-1,j})}) + l(v_{i-1,j})$;

$\Gamma(v_{i-1,j}) = \Gamma(v_{i,\tau(v_{i-1,j})}) \bigcup \{\tau(v_{i-1,j})\}$;

$i = i-1$;

终止： $L^*_{bottom-half} = \min_{x^n - r_{THS} \le j \le x^n + r_{THS}} L(v_{mid^n_{bottom},j})$;

$j^* = \arg\min_{x^n - r_{THS} \le j \le x^n + r_{THS}} L(v_{mid^n_{bottom},j})$;

路径求取： $\Gamma^*_{bottom-half} = \Gamma(v_{mid^n_{bottom},j^*})$.

图 3 切分区 R^n 切分路径搜索的 Viterbi 算法

Fig. 3 Viterbi search algorithm for character segmentation path in the R^n region

为了利用切分区中各象素点之间的关系和相关的二值及字符轮廓信息，我们采用 Viterbi 算法来进行上述路径的搜索. 设 $l(v_{i,k},v_{i+1,j})$ 为边 $<v_{i,k},v_{i+1,j}>$ 的路径长度，在我们的系统中，有

$$l(v_{i,k},v_{i+1,j})=\begin{cases}-[\log\frac{(1/5+\alpha_{i,k}+\beta_{i,k})}{\sum_{m=j-2}^{j+2}(1/5+\alpha_{i,k}+\beta_{i,k})}+\log\frac{p(i,k)}{255}] & j-2\le k\le j+2\\ \infty & \text{otherwise}\end{cases}\tag{11}$$

其中，$p(i,k)$ 为该点的灰度值；$0\le\alpha_{i,k}<1$ 表示如果 $v_{i,k}$ 对应的为字符轮廓点时路径的加权参数，$0\le\beta_{i,k}<\alpha_{i,k}$ 表示对应的二值图像为黑点时的加权参数.

在这里，$\frac{p(i,k)}{255}$ 可以看作点 $v_{i,k}$ 的观察值 $p(i,k)$ 所属某一状态的发生概率；而表达式 $(1/5+\alpha_{i,k}+\beta_{i,k})/\sum_{m=j-2}^{j+2}(1/5+\alpha_{i,k}+\beta_{i,k})$ 可以看作点 $v_{i,k}$ 的状态到点 $v_{i+1,j}$ 的状态的转移概率.

此时，在切分区 R^n 基于 Viterbi 算法的非线性切分路径搜索如图 3 所示. 其中，$L(v_{i+1,j})$ 表示点 $v_{i+1,j}$ 处的累计路径长度，$\tau(v_{i+1,j})$ 表示在点 $v_{i+1,j}$ 处 Viterbi 算法所得到的有效结点 (Survivor)，而 $\Gamma(v_{i+1,j})$ 表示到点 $v_{i+1,j}$ 处的路径所有 Survivor 点的集合.

最后，根据 δ_n 合并两次 Viterbi 搜索的路径. 如果 $\delta_n<\delta_{THS}$，则直接合并上述两次 Viterbi 搜索的路径，有

$$\Gamma*=\Gamma^*_{top-half}\cup\Gamma^*_{bottom-half}\tag{12}$$

如果 $\delta_n\ge\delta_{THS}$，则需要增加中间区域的路径. 对于密度大的情况，在该中间区域采用直线切分路径即可，设其为 Γ^*_{middle}. 此时，最终的切分路径为

$$\Gamma*=\Gamma^*_{top-half}\cup\Gamma^*_{middle}\cup\Gamma^*_{bottom-half}\tag{13}$$

4.2　基于 Viterbi 算法的切分与识别

基于 Viterbi 算法的非线性切分路径搜索，解决了字符行粘连的情况. 由于粘连断裂的存在，上述的初始切分会带来多余的切分点，即所谓的 “过切分” (Over segmentation). 所以，需要一定的策略对这些切分点进行合并处理. 很多研究者把待处理的字符行建立成一个基于切分点的图结构，然后利用动态规划方法进行切分识别 [2,3,13]；其中，[13] 还利用一种限制型 Viterbi 算法来有效的选择识别器的正反训练样本. 在我们的系统中，把由候选切分点得到的切分块看作字符状态观测序列，利用 Viterbi 算法求相关的最短路径 (最大可信路径)，对粘连断裂印刷体数字行进行切分识别.

针对候选切分点序列$\{x^1,x^2,\ldots,x^N\}$，我们假定，最多有相邻的 4 个切分点组成一个合理的切分块，最少为两个切分点.

第 1 个切分点为起始点. 一般的，对于切分点 $i\,(i>3)$，共可组成三个切分块：$Seg(i-3,i)$、$Seg(i-2,i)$ 和 $Seg(i-1,i)$. 为了方便的使用 Viterbi 算法，在切分点 i 处，我们还设立了两个空的切分块，$Seg(-1,i)$ 表示候选切分点 i 处不看作真正的切分点，即切分从 $i-1$ 处直接跳跃至 $i+1$ 处；$Seg(-2,i)$ 表示候选切分点 i 处不看作真正的切分点，而且 $i-1$ 处也不看作真正的切分点，即切分从 $i-2$ 处直接跳跃至 $i+1$ 处.

如果把每一个切分块 (包括空切分块) 看作一个结点，则可能的切分情况如图 4 所示. 我们的目的就是，从第一个切分点 1 开始，到最后一个切分点 N 结束，寻找一条最短的切分路径. 图 4 中，我们以 $i=5$ 来进行说明，虚线 (不可达路径) 表示这前后切分点是不可达的；而实线 (可达路径) 表示在第 5 切分点某一种切分情况时，其在第 4 切分点处几种可能的切分.

在切分点 $i+1$ 处，利用前面介绍的分类器，对三个非空切分块分别进行识别，得到三个识别结果，设为 $Lab((i+1)-3,i+1)$、$Lab((i+1)-2,i+1)$ 和 $Lab((i+1)-1,i+1)$，其相应的可信度分别为 $S(i+1)-3,i+1)$、$S((i+1)-2,i+1)$ 和 $S((i+1)-1,i+1)$. 对于空切分情况，虽然没有实际的识别，但是为了描述方便，仍以 $Lab(-1,i+1)$、$Lab(-2,i+1)$、$S(-1,i+1)$、$S(-2,i+1)$ 来表示其识别结果和可信度. 对于切分点 $i+1$ 处的第 j 种切分，从切分点 i 处第 m 种切分到达的路径长度为 (可达路径)

$$l(O^j_{i+1},O^m_i)=\begin{cases}0 & m<0\\ -[\log P(Lab(j,i+1),Lab(m,i))+\log S(m,i)] & \text{otherwise}\end{cases}\tag{14}$$

其中，$P(Lab(j,i+1),Lab(m,i))$ 为状态 $Lab(m,i)$ 到状态 $Lab(j,i+1)$ 的转移概率，这里有

$$P(Lab(j,i+1),Lab(m,i)=1/5,\quad (\ i>1\ )\tag{15}$$

对于不可达路径，有 $l(O^j_{i+1},O^m_i)=\infty$.

本系统提出的一阶 Viterbi 切分识别的方法如图 5 所示. 其中，$L(v_{i+1,j})$ 表示点 $v_{i+1,j}$ 处的累计路径长度，$\tau(v_{i+1,j})$ 表示在点 $v_{i+1,j}$ 处 Viterbi 算法所得到的 Survivor 结点，而 $\Gamma(v_{i+1,j})$ 表示到点 $v_{i+1,j}$ 处路径所包括 Survivor 点的集合.

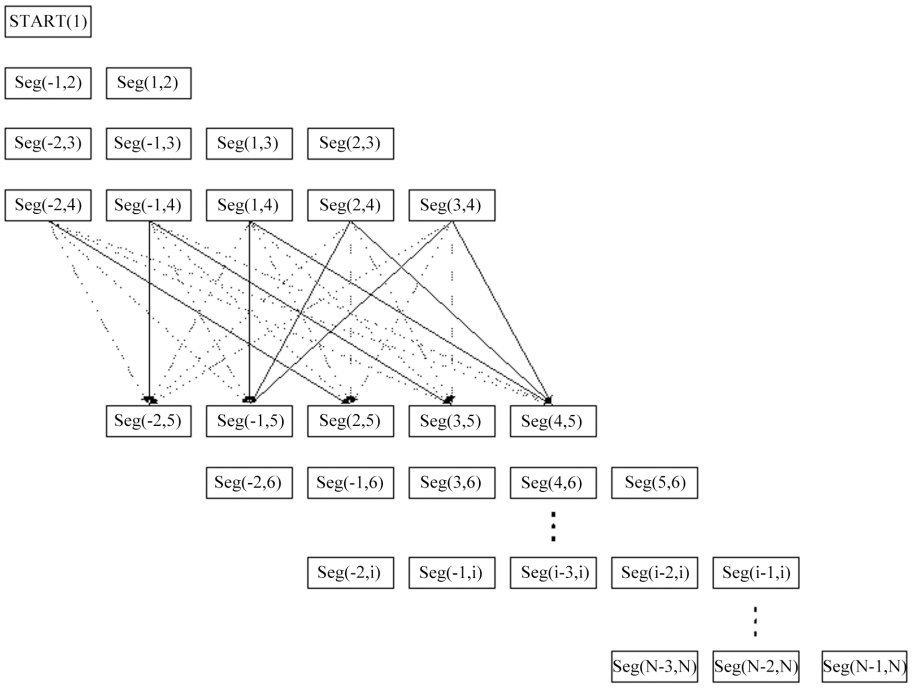

图 4 N 个候选切分点的切分情况图

Fig. 4 Segmentation paths for N candidate points

初始化：$i=1$； $L(1)=0$； $\Gamma(1)=\{\ \}$；

递归： For $-2\le k_{i+1}\le 2$

$$j=\begin{cases}k_{i+1}, & k_i<0\\ (i+1)-k_i-1, & otherwise\end{cases};\quad j=\max(j,(i+1)-1);$$

For $-2\le k_i\le 2$

$$m=\begin{cases}k_i, & k_{i-1}<0\\ i-k_{i-1}-1, & otherwise\end{cases};\quad m=\max(m,i-1);$$

If 不可达路径

$$l(O_{i+1}^{j},O_i^{m})=\infty;$$

Else If $m<0$

$$l(O_{i+1}^{j},O_i^{m})=0;$$

Else

$$l(O_{i+1}^{j},O_i^{m})=-[\log P(Lab(j,i+1),Lab(m,i))+\log S(m,i)];$$

$$l(j,i+1)=\min_m l(O_{i+1}^{j},O_i^{m});$$

$$\tau(j,i+1)=\arg\min_m l(O_{i+1}^{j},O_i^{m});$$

$$L(j,i+1)=L(\tau(j,i+1),i)+l(j,i+1);$$

$$\Gamma(j,i+1)=\Gamma(\tau(j,i+1),i)\cup\{\tau(j,i+1)\};$$

$$i=i+1;$$

终止： $L^*=\min_j L(j,N)$； $j^*=\arg\min_j L(j,N)$；

路径求取：$\Gamma^*=\Gamma(j^*,N)$.

图 5 基于 Viterbi 算法的切分与识别

Fig. 5 Segmentation and recognition based on Viterbi

求得最大可信切分识别路径后，就很容易知道相应的识别结果了. 如果路径中存在空的切分结点，则在结果输出时，跳过该结点即可.

该切分识别方法存在的一个问题就是，如果该字符行包括的候选切分点过多，其切分识别效率是比较低的. 在实际应用中，该问题可以根据规则知识来克服. 在这里，我们也应用如下规则，来提高处理的速度 (其中 W_{Max} 和 W_{Min} 分别为事先选定的字符最大和最小宽度)：

1) 在“候选切分点定位”过程中，如果某个切分点是非常可信的 (例如该处没有粘连与断裂情况)，则可以把所处理的切分路径分割成两条，每条路径分别独立的利用上述 Viterbi 算法进行切分与识别；

2) 在运行 Viterbi 算法的过程中，如果在第 i 步，有$\mathrm{Seg}(i-3,i)$ $>\mathrm{W}_{Max}$，则直接赋值 $l(O_i^j,O_{i-1}^m)=\infty$，而不需要用分类器进行识别，对于$\mathrm{Seg}(i-2,i)$也有同样的处理；

3) 在运行 Viterbi 算法的过程中，如果在第 i 步，有$\mathrm{Seg}(i-1,i)$ $<\mathrm{W}_{Min}$，则直接赋值 $l(O_i^j,O_{i-1}^m)=0$，也不需要用分类器进行识别.

5 实验与分析

本实验是在交通银行金融票据识别系统的基础上进行的，测试的对象为票据上的打印体 (印刷体) 流水号数字行. 这些字符行由于打印机的不同、打

印字体的不同和打印油墨的浓淡不同，往往产生复杂的粘连与断裂情况. 通过票据分类与区域提取系统[14]，我们收集了南宁交行和兰州交行某一整天所有票据上的印刷体字符行图像块样本 (南宁交行的样本数为 2305 块，兰州交行的样本数为 2978 块).

为了分析本文提出的识别系统对粘连断裂字符行切分识别的效果，对于这些测试样本，采用三种方法进行切分识别：方法 1 是该系统中的第一次切分识别；方法 2 类似于方法 1，只是利用了 4.1.2 节 “ 基于 Viterbi 算法的非线性切分路径搜索” 来处理第一次切分识别过程中切分点处的切分路径搜索，以便于观察非直线切分路径对粘连字符切分的影响；方法 3 就是本文提出的基于 Viterbi 算法的粘连断裂印刷体数字行切分识别方法. 三种方法对于上述样本数据的测试结果如表 1 所示. 其中 “正确” 表示字符行识别正确 (字符行中所有的字符都识别正确) 的块数，“识别率” 表示 “正确” 数与总块数的比值.

表 1　数字行切分识别测试结果

Table 1　Machine-printed numeral string recognition results

	南宁交行样本		兰州交行样本		累计	
	正确	识别率	正确	识别率	正确	识别率
方法 1	2207	95.75%	2742	92.08%	4949	93.68%
方法 2	2214	96.05%	2747	92.24%	4961	93.90%
方法 3	2253	97.74%	2777	93.25%	5030	95.21%
样本数	2305		2978		5283	

通过表 1 可以看出，方法 2 中使用的非直线切分路径能够在一定程度上克服粘连字符的切分问题；而且我们发现，方法 2 比方法 1 多识别正确的样本都是粘连字符行. 在图 6 第一行图像中，第一个图为方法 1 得到的直线切分路径，由于 “03” 粘连在一起，其结果识别为 “08”，第三个图也是方法 1 得到的直线切分路径，由于 “26” 粘连在一起，其结果识别为 “25”；第二个图和第四个图为方法 2 的相应非直线切分路径，这表明基于 Viterbi 算法的非线性切分路径搜索能够较好的对粘连字符进行切分. 但是，还有较多的粘连字符行不能被方法 2 正确识别，其主要原因是在第一次切分识别过程中，由于字符粘连情况比较复杂，初始候选切分点定位不准确，从而导致识别失败；另外对于断裂字符行，方法 2 就显得无能为力. 方法 3 较好的解决了这些问题. 在图 6 第二行图像中，第一个图和第三个图为方法 2 得到的切分路径和识别结果，由于 “55” 和 “05” 粘连严重，难以准确切分，而第二个图和第四个图为方法 3 的相应切分识别结果，能够较好的对粘连字符进行切分识别；第三行图像中第一个图和第三个图为方法 2 的切分识别结果，由于字符断裂情况严重，导致切分位置不准确，而第二和第四个图为方法 3 的结果，首先进行 “过切分”，然后利用 Viterbi 算法进行合并.

图 6 三种方法切分路径与识别结果比较

Fig. 6　Segmentation paths and recognition results of three experimental methods

6　结论

针对金融票据识别系统中存在的大量粘连断裂字符情况，本文提出了一种基于 Viterbi 算法的印刷体数字行切分识别系统. 在我们的识别系统中，采用两次切分识别过程. 第一次切分识别直接在二值图像上进行. 如果第一次切分识别的可信度都比较高，则整个识别系统结束；否则进行第二次切分识别处理. 第二次切分识别结合灰度和二值信息进行，首先利用 Viterbi 算法，结合二值及轮廓信息在灰度图像上搜索定位非直线的切分路径；然后再次利用 Viterbi 算法，在识别可信度的基础上动态合并这些非直线的切分路径进行切分识别，得到最终的字符行识别结果.

相比与普通的动态规划方法[2,3]，不管是切分路径的搜索，还是切分候选块的动态合并切分识别，Viterbi 算法能够较好的描述路径搜索和切分识别中的动态特性. 而且，对于先验知识与规则，可以方便的应用于 Viterbi 算法中. 实验表明，本文提出的基于 Viterbi 算法的粘连断裂印刷体数字行切分识别方法能够较好的克服字符行的粘连与断裂情况，对于提高切分识别系统的识别率和鲁棒性都有较好的效果.

致谢

感谢邹明福博士、娄正良博士和汉王科技研发中心 OCR 软件部各位工程师的帮助. 同时也感谢南开大学软件学院韩智博士的有益讨论.

References

1　Nagy G. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(1): 38∼62

2 Lee S W, Lee D J, Park H S. A new methodology for gray-scale character segmentation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1996, **18**(10): 1045∼1050

3 Arica N, Yarman-Vural F T. Optical character recognition for cursive handwriting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2002, **24**(6): 801∼813

4 Casey R G, Lecolinet E. A survey of methods and strategies in character segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, **18**(7): 690∼706

5 Forney G D. The Viterbi algorithm. *Proceedings of the IEEE*. 1973, **61**(3): 268∼278

6 Viterbi A J. Error bounds for convolution codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 1967, IT-**13**(4): 260∼269

7 Rabiner L R. A Tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989, **77**(2): 257∼286

8 Shinghal R, Toussaint G T. Experiments in text recognition with the modified Viterbi algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979, **1**(2): 184∼192

9 Lu Y. On the segmentation of touching characters. In: Proceedings of International Conference on Document Analysis and Recognition. New Jersey, 1993. 440∼443

10 Sankur B, Sezgin M. A survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*. 2004, **13**(1): 146∼165

11 Otsu N. A threshold selection method from gray level histograms. *Pattern Recognition*, 1979, **9**(1): 62∼66

12 Niblack W. *An Introduction to Image Processing*. New Jersey: Prentice-Hall, 1986. 115∼116

13 Burges C J C, Matan O, LeCun Y, *et al*. Shortest path segmentation: A method for training a neural network to recognize character strings. In: Proceedings of International Joint Conference on Neural Networks, New Jersey, 1992. **3**: 165∼172

14 Yin Xu-Cheng , Jiang Shi-Sheng, Han Zhi, Liu Chang-Ping. A hierarchical method for form classification of financial document images. *Journal of Chinese Information Processing*, 2005, **19**(6): 70∼77
(殷绪成, 江世盛, 韩智, 刘昌平. 层次型金融票据图像分类方法. 中文信息学报, 2005, **19**(6): 70∼77)

殷绪成 2006 年毕业于中国科学院自动化研究所，获博士学位，现工作于富士通研究开发中心有限公司，其主要研究领域为模式识别、机器学习、计算机视觉和图像处理. 本文通信作者. E-mail: xuchengyin@hotmail.com
(**YIN Xu-Cheng** Ph.D. Graduated from Institute of Automation, Chinese Academy of Sciences on July 2006. Currently, he works at Fujitsu R&D Center Co., LTD. His research interest covers pattern recognition, machine learning, computer vision, and image processing. Corresponding author of this paper.)

刘昌平 中国科学院自动化研究所研究员，主要研究领域为：模式识别、字符识别和人机交互技术等. E-mail: changping.liu@ia.ac.cn
(**LIU Chang-Ping** Professor of Institute of Automation, Chinese Academy of Sciences. His research interest covers pattern recognition, OCR, and HCI.)