一种视频中字符的集成型切分与识别算法

杨武夷 [1, 2] 张树武 [3]

摘 要 视频文本行图像识别的技术难点主要来源于两个方面: 1) 粘连字符的切分与识别问题; 2) 复杂背景中字符的切分与识别问题. 为了能够同时切分和识别这两种情况中的字符, 提出了一种集成型的字符切分与识别算法. 该集成型算法首先对文本行图像二值化, 基于二值化的文本行图像的水平投影估计文本行高度. 其次根据字符笔划粘连的程度, 基于图像分析或字符识别对二值图像中的宽连通域进行切分. 然后基于字符识别组合连通域得到候选识别结果, 最后根据候选识别结果构造词图, 基于语言模型从词图中选出字符识别结果. 实验表明该集成型算法大大降低了粘连字符及复杂背景中字符的识别错误率.

关键词 视频文字, 集成切分与识别, 字符切分, 字符识别

DOI 10.3724/SP.J.1004.2010.01468

An Integrated Segmentation and Recognition Algorithm for Text in Video

YANG Wu-Yi[1, 2] ZHANG Shu-Wu[3]

Abstract There are two difficulties to recognize the text images which are extracted from videos: 1) how to segment and recognize the merged characters; 2) how to segment and recognize the characters with complex backgrounds. To overcome the difficulties, a novel integrated segmentation and recognition method is proposed. The method first binarizes the text image and estimates the height of the text line. Second, the connected components in the binary text image, which are wider than a threshold, are segmented based on image analysis or character recognition. Third, the connected components are selected and combined to generate the character patterns based on character recognition. Last, the best character sequence is selected based on a statistical language model. Experimental results demonstrate the effectiveness of the proposed method.

Key words Video text, integrated segmentation and recognition, character segmentation, character recognition

视频文字直接承载了高层语义信息. 因此, 如果能够有效地提取视频中的文字信息, 对高速增长的视频内容的高效检索、理解和复用具有重要的作用[1]. 从视频中提取的文本行图像种类繁多, 字符识别的技术难点主要来源于两个方面: 一方面, 由于视频的数据压缩, 视频中的文字出现相邻笔划粘连等退化现象, 如图 1 (a) 所示; 另一方面, 视频中的文字可能存在于剧烈变化的复杂背景中, 如图 1 (b) 所示. 对于剧烈变化的复杂背景, 目前的文字分割方法还得不到理想的二值图像, 二值图像中很可能存在非字符的 "噪声" 成分. 目前, 研究人员针对视频中文字的定位和分割问题进行了较深入的研究. 对于字符识别, 已有的视频文字提取系统通常把文字分割后得到的二值图像输入传统的识别引擎进行字符识别[1−2].

(a) 字符笔划粘连的文本行图像

(a) Text image with merged character strokes

(b) 背景复杂的文本行图像

(b) Text image with complex background

图 1 文本行图像

Fig. 1 Text image

目前的识别算法主要分为串行的字符切分与识别算法[3−6] 和集成型的字符切分与识别算法[7−12]. 串行的字符切分与识别算法基于图像分析对文本行图像进行切分, 然后对切分单元进行字符识别. 基于图像分析的方法一般在二值化的文本行图像上进行, 其又可分为基于垂直投影的方法[3−5] 和基于连通域分析的方法[6]. 投影法的优点是速度快, 算法简单; 缺点是当字符粘连较严重或者字符上下交叠较严重时, 不能准确地判断切分位置. 基于连通域分析的方法能将上下交叠的非粘连字符分开, 但不能分开粘连字符. 串行的字符切分与识别算法没有形成有效的反馈, 字符切分过程无法利用识别的信息, 导致一些复杂情况中的字符不能得到准确的切分与识别.

收稿日期 2009-10-30 录用日期 2010-06-03
Manuscript received October 30, 2009; accepted June 3, 2010

1. 水声通信与海洋信息技术教育部重点实验室 (厦门大学) 厦门 361005 2. 厦门大学海洋与环境学院 厦门 361005 3. 中国科学院自动化研究所 北京 100190

1. Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Xiamen University), Ministry of Education, Xiamen 361005 2. College of Oceanography and Environmental Science, Xiamen University, Xiamen 361005 3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190

集成型的字符切分与识别算法充分利用识别的信息. 为了解决粘连字符的切分与识别问题, 文献[7−8] 提出了基于滑动窗方法. 文献 [9] 首先将文本行图像切分成小的切分单元, 然后构造候选字符切分图, 最后基于候选切分单元的字符识别得分得到最优的字符识别结果. 文献 [10] 根据从文本行图像中提取的特征推测可能的候选字符, 然后基于相应的字模把字符切分出来, 产生候选的字符切分和识别结果, 在各种字符切分的基础上对剩余的文本行图像进行类似的处理, 直至文本行图像切分完毕, 得到各种切分方案, 最后选出最优的切分方案和字符识别结果. 这些集成型算法主要是针对文档扫描图像的字符切分与识别问题, 能较好地对粘连字符进行切分与识别, 但不能有效地处理包含 “噪声” 连通域的文本行图像.

前面介绍的方法不能很好地处理视频中字符的切分和识别问题. 为了能够克服视频文本行图像识别的技术难点, 本文提出了一种集成型字符切分与识别处理算法, 算法的总框架如图 2 所示. 算法首先对文本行图像二值化, 基于二值图像的水平投影估计文本行的高度. 然后对二值图像进行连通域分析, 切分宽度较大的连通域. 为了克服二值图像中非字符 “噪声” 成分对字符切分与识别的影响, 遍历连通域的各种组合, 基于字符识别组合连通域得到候选识别结果. 最后根据候选识别结果构造词图, 基于统计语言模型从词图中选出字符识别结果.

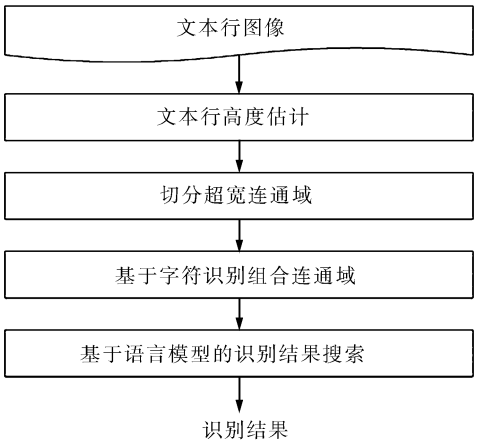


图 2　算法总框架

Fig. 2　Overview of the proposed method

本文的方法与文献 [11] 中的方法类似. 文献 [11] 针对手写文本行图像的识别, 提供了一个简洁的切分 – 识别框架, 但是候选切分 – 识别的搜索空间会比较大. 由于处理的数据对象不同, 本文的方法与文献 [11] 中的方法在具体实现上有较大区别, 特别是在构造词图之前, 本文的方法有效削减了候选切分和识别结果的个数.

1　文本行高度估计

1.1　文本行图像二值化

本文利用文献 [13] 的方法对文本行图像二值化, 算法首先从不同的角度提取图像的信息, 得到不同的二值图像, 然后把这些二值图像融合, 得到最终的二值图像. 设得到的二值图像为 $B_0 = [b_0(x,y)]$, 其中位于第 y 行第 x 列的像素值为 $b_0(x,y)$, $b_0(x,y)=0$ 代表背景像素, 1 代表字符笔划像素.

1.2　文本行高度估计

设二值图像 B_0 的宽度和高度分别为 w 和 h, 其在水平方向上的投影 $P_h(y)$ 定义为

$$P_h(y)=\sum_{x=0}^{w-1} b_0(x,y),\quad y=0,\cdots,h-1 \tag{1}$$

利用 $L(y)$ 表示第 y 行是否包含字符, $L(y)$ 被定义为

$$L(y)=\begin{cases}0, & \text{若 } P_h(y)<T\\ 1, & \text{其他}\end{cases},\quad y=0,\cdots,h-1 \tag{2}$$

其中, $T=a\cdot w$, a 取经验值 0.08. 设 $L(y)$ 中最长的 1 游程的起始和结束位置分别为 tpy 和 bty, 则文本行高度 $H=bty-tpy+1$.

定义二值图像 $B=[b(x,y)]$ 为

$$b(x,y)=\begin{cases}b_0(x,y), & \text{若 } tpy\le y\le bty\\ 0, & \text{其他}\end{cases} \tag{3}$$

经过上述处理得到的二值图像 $B=[b(x,y)]$ 可以去除文本行外一些误判为字符笔划像素的背景像素. 下面将利用二值图像 B 进行字符的集成切分与识别处理.

2　切分超宽连通域

由于数据压缩作用, 视频文字会出现边缘模糊、笔划断裂、相邻字符笔划合并等退化现象. 相邻字符笔划合并导致相邻的字符组成一个连通域, 在二值图像中表现为一些较宽的连通域, 如图 3 所示. 连通域切分的目的就是要把由多个字符组成的连通域切分成宽度较小的多个连通域, 每个连通域对应单独的一个字符或字符的某个部分, 为基于字符识别组合连通域的处理做好准备.

2.1　连通域分析

对二值图像 B 进行连通域分析, 得到每个连通域的编号、像素个数和位置信息. 对图 1 (a) 进行连通域分析的结果如图 3 所示, 图中每个连通域用矩形

图 3 二值图像中的连通域

Fig. 3 Connected components in the binary image

外框框住了, 可以发现图中相邻的一些字符合并为一个连通域.

设在二值图像中共找到 N 个连通域, 第 i 个连通域 $C(i)$ 的矩形外框的宽度为 $w(i)$, $l(i)$ 为矩形外框左边界在图像 X 轴方向上的位置, $r(i)$ 为矩形外框右边界在图像 X 轴方向上的位置, 则 $r(i) = l(i) + w(i) - 1$. 如果第 i 个连通域 $C(i)$ 和第 j 个连通域 $C(j)$ 满足

$$\frac{\min(r(i), r(j)) - \max(l(i), l(j)) + 1}{\min(w(i), w(j))} = 1 \quad (4)$$

则合并这两个连通域. 重复上述过程直至两两连通域不再满足上面的条件为止.

2.2 切分超宽连通域

判断一个连通域是否需要对其进行切分是以其宽度是否大于 $r_m \times H$ 为基准, r_m 为字符的最大宽高比, H 为估计的文本行高度, 取经验值 $r_m = 1.2$. 对一个超宽连通域, 首先基于二值图像垂直投影对其进行切分, 如果切分处理失败, 则利用基于滑动窗口的方法对超宽连通域进行切分. 重复上述过程直至图像中所有的连通域的宽度都不大于 $r_m \times H$.

2.3 基于二值图像垂直投影切分连通域

利用二值图像在垂直方向的投影 $P_v(x)$ 寻找竖直的切分线对超宽连通域 C 进行切分. 为了在字符粘连的位置产生更小的投影值, 首先对二值图像 $b(x, y)$ 中相邻的两列像素进行 "与" 运算[14], 得到二值图像 $H = [h(x, y)]$. 二值图像 $H = [h(x, y)]$ 定义为

$$h(x, y) = b(x, y) \ \& \ b(x + 1, y), \quad x = 0, \cdots, w - 2, \ \ y = 0, \cdots, h - 1 \quad (5)$$

求二值图像 $h(x, y)$ 在垂直方向的投影 $P_v(x)$

$$P_v(x) = \sum_{y=0}^{h-1} h(x, y), \ \ x = 0, \cdots, w - 1 \quad (6)$$

设连通域 C 的矩形外框左上角在二值图像 B 中的位置为 (lx, ty). 令 $begx = \text{int}(lx + 0.25H)$, $endx = \text{int}(lx + 1.2H)$, 其中函数 $\text{int}(x)$ 定义为求小于等于 x 的最大整数. 在 $begx$ 至 $endx$ 的范围内寻找 $P_v(x)$ 的最小值 P_m, 并计算这个范围内垂直投影的均值 P_a:

$$P_m = \min\{P_v(x), \ \ x = begx, \cdots, endx\} \quad (7)$$

$$P_a = \frac{\sum_{x=begx}^{endx} P_v(x)}{endx - begx + 1} \quad (8)$$

设最小投影值 P_m 出现的位置为 p, 即 $P_m = P_v(p)$. 如果在 $begx$ 至 $endx$ 的范围内, 有多个位置的投影值等于最小值 P_m, 则在 $begx$ 至 $lx + H$ 的范围内寻找投影值为 P_m 的位置, 把此范围内最右边那个投影值为 P_m 的位置作为 p 的值. 如果最小投影值 $P_m \le P_a/b$, 则表示字符粘连不严重, 将 p 作为切分线的位置对连通域 C 进行切分, 判定基于投影的切分处理成功; 如果 $P_m > P_a/b$, 垂直投影在 p 处没有明显的低谷, 则 p 不一定是准确的切分位置, 判定基于投影的切分处理失败. b 取经验值 2.

2.4 基于字符识别切分连通域

当字符之间粘连比较严重, 为了得到可靠的切分位置, 将充分利用字符识别器. 将一个宽度为 H 的窗口滑过待切分的连通域, 滑动步长为 1, 对窗口内的图像进行字符识别, 得到窗口内图像是字符的置信度. 超宽连通域 C, 其矩形外框左边界在图像 X 轴方向上的位置为 lx, 则滑动窗口左边界的滑动范围为 $[begx, endx]$, $begx = \text{int}(lx - 0.6H)$, $endx = \text{int}(lx + 0.6H)$. 找到置信度最大的窗口, 把窗口的左右边界作为垂直切分线的位置对连通域进行切分. 滑动窗口左边界的滑动范围不从 lx 开始是因为超宽连通域 C 的左边部分可能是与其左边的连通域构成一个字符, 而且其置信度可能比较高, 因此可能可以找到更准确的切分位置.

重复上述过程, 对图 3 中的超宽连通域进行切分, 直至图像中所有的连通域的宽度都不大于 1.2 倍的文本行高度, 得到的切分结果如图 4 所示.

图 4 切分处理后的连通域

Fig. 4 Connected components after the cutting process

3 基于字符识别组合连通域

字符可能由多个连通域组成, 同时字符之间或内部可能还包含 "噪声" 连通域, 如图 5 所示.

图 5 包含 "噪声" 连通域的二值图像

Fig. 5 Binary image contains "noise" components

为了使字符识别不受 “噪声” 连通域的影响, 本文将基于字符识别组合连通域, 得到各种候选的识别结果.

3.1 判断连通域或连通域的组合是否可能为字符

连通域或连通域的组合 C 的矩形外框在图像中的位置用 $R((lx,ty),w,h)$ 表示, 其中, (lx,ty) 为矩形外框左上角在图像中的坐标, w 和 h 分别为外框的宽和高. C 的矩形外框右下角在图像中的坐标为 (rx,by), $rx=lx+w-1$, $by=ty+h-1$. 令 $W_{\max}=r_m\times H$ 为字符可能的最大宽度, 其中 r_m 为字符的最大宽高比, 取经验值 $r_m=1.2$. $W_{\min}=r_c\times H$ 为中文字符的最小宽度, 其中 r_c 为中文字符的最小宽高比, 取经验值 $r_c=0.7$. tpy 和 bty 为第 1.2 节中求得的文本起始行和文本终止行在 Y 轴方向上的位置. 令 $maxty=tpy+a\times H$ 为中文字符矩形外框的左上角在 Y 轴方向上的最大坐标值, $minby=bty-a\times H$ 为中文字符矩形外框的右下角在 Y 轴方向上的最小坐标值, a 取经验值 0.2. 对 C 进行字符识别, C 与候选字符类距离的最小值为 $D(C)$. 如果 $D(C)$ 小于等于阈值 α, 则判断 C 可能是一个字符; 否则判断 C 不是一个字符. C 为一个字符的置信度定义为 $-D(C)$. 当 $D(C)$ 小于等于阈值 α, 我们利用文献 [15] 中的方法确定 C 的候选字符类数, 该方法把所有与 C 的距离小于等于 $D(C)\times\beta$ 且小于等于 α 的字符类当成 C 的候选字符类.

定义处理子过程, 判断一个连通域或连通域的组合 C 是否可能为一个字符. 子过程返回处理成功或失败分别代表 C 可能是一个字符或不是一个字符, 其处理流程如下:

步骤 1. 如果 C 的矩形外框位置满足:

1) $maxty<ty$ 且 $by<minby$;

2) $W_{\min}<w<W_{\max}$;

3) $h<2.5w$;

4) $D(C)\le\alpha$, 则返回处理成功, 否则进入步骤 2.

步骤 2. 如果 C 的矩形外框位置满足:

1) $maxty>ty$ 且 $by>minby$;

2) $w<W_{\min}$;

3) $D(C)\le\alpha$ 且 C 的字符识别结果为数字或标点符号, 则返回处理成功, 否则进入步骤 3.

步骤 3. 如果 C 的矩形外框位置满足:

1) $maxty>ty$ 且 $by>minby$;

2) $W_{\min}<w<W_{\max}$;

3) $D(C)\le\alpha$, 则返回处理成功, 否则返回处理失败.

上述步骤 1 是为了判断 C 是否为字符 “一”; 步骤 2 是为了判断 C 是否为数字或标点符号; 步骤 3 是为了判断 C 是否为中文字符.

3.2 基于字符识别组合连通域

设连通域切分处理后, 图像中共有 M 个连通域, 第 i 个连通域为 $C(i)$, $C(i)$ 的矩形外框在图像中的位置用 $R((lx(i),ty(i)),w(i),h(i))$ 表示, 其中 $(lx(i),\ ty(i))$ 为矩形外框的左上角在图像中的坐标, $w(i)$ 和 $h(i)$ 分别为矩形外框的宽和高. $C(i)$ 的矩形外框右下角在图像中的坐标为 $(rx(i),by(i))$, $rx(i)=lx(i)+w(i)-1$, $by(i)=ty(i)+h(i)-1$. 把可能为字符的连通域或者连通域组合的相关信息保存在队列 $RCcList$ 中, 基于字符识别组合连通域的处理流程如下:

步骤 1. 初始化: 把全部连通域放入排序链表 $CcOrdList$ 中, 依据连通域矩形外框左上角在 X 轴方向上的坐标值, 从左到右对连通域进行排序.

步骤 2. 从 $CcOrdList$ 中取出第一个连通域 C, 对连通域 C 调用上述判断一个连通域是否可能为字符的处理子过程, 如果返回处理成功, 则把 C 的位置、C 的候选字符类、C 与候选字符类的距离及 C 为一个字符的置信度保存在候选结果队列 $RCcList$ 中.

步骤 3. 参考连通域 C 的位置, 遍历 $CcOrdList$ 中的所有连通域, 如果 $CcOrdList$ 中的第 k 个连通域满足 $lx(k)\ge lx$ 且 $rx(k)-lx<W_{\max}$, 其中 lx 为 C 的矩形外框左上角在 X 轴上的坐标, 则把第 k 个连通域保存在队列 $CcCadList$ 中.

步骤 4. 如果队列 $CcCadList$ 中连通域的个数为 0, 则进入步骤 6 进行处理, 否则进入步骤 5 处理 $CcCadList$ 中的连通域.

步骤 5. 遍历队列 $CcCadList$ 中连通域所有可能的组合, 每种连通域的组合与连通域 C 再组合成连通域组合 CA, 然后对 CA 调用判断连通域组合是否可能为字符的处理子过程, 如果返回处理成功, 则把 CA 的位置、CA 的候选字符类、CA 与候选字符类的距离及 CA 为一个字符的置信度保存在候选结果队列 $RCcList$ 中; 然后进入步骤 6 进行处理.

步骤 6. 如果 $CcOrdList$ 为非空, 则进入步骤 2 继续处理; 否则退出基于字符识别组合连通域的处理.

基于字符识别组合图 5 中的连通域得到的结果如图 6 所示, 对可能是字符的连通域或连通域的组合, 其字符识别的候选结果在连通域或连通域组合的下面. 虚线框中的字符为同一个连通域或者连通域组合的候选识别结果.

4 基于统计语言模型的识别结果搜索

为了叙述方便, 下面将队列 $RCcList$ 中的连通

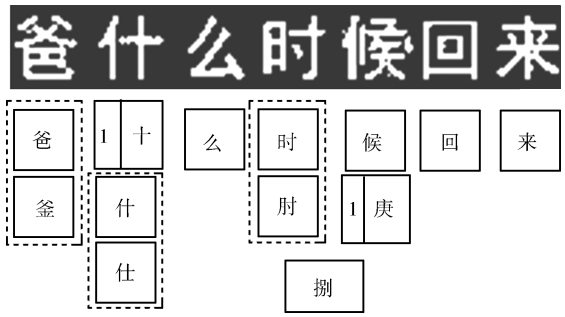


图 6 基于字符识别组合连通域后的处理结果

Fig. 6 Results by assembling components based on character recognition

域和连通域的组合统称为切分单元. 基于语言模型的识别结果搜索从切分单元的候选识别结果中选出最终的识别结果.

设文本行图像的一种切分方法为 $\mathbb{C} = C_1C_2\cdots C_l = c_1^1\cdots c_1^{m_1}c_2^1\cdots c_2^{m_2}\cdots c_l^1\cdots c_l^{m_l}$, $\mathbb{C}$ 被识别成词序列 $\mathbb{W} = W_1W_2\cdots W_l$, 词序列 $\mathbb{W}$ 中的第 k 个词 W_k 由 m_k 个切分单元 $C_k = c_k^1\cdots c_k^{m_k}$ 的候选识别结果组成. 基于统计语言模型的识别结果搜索就是寻找词序列 $\hat{\mathbb{W}}$ 使得

$$\hat{\mathbb{W}} = \arg\max_{\mathbb{W}} P(\mathbb{W}|\mathbb{C}) \tag{9}$$

根据 Bayes 定理

$$\hat{\mathbb{W}} = \arg\max_{\mathbb{W}} P(\mathbb{C}|\mathbb{W})P(\mathbb{W}) \tag{10}$$

对于 N 元语言模型, 词序列 $\mathbb{W}$ 的概率为[16]

$$P(\mathbb{W}) = \prod_{i=1}^{l} P(W_i|W_{i-N+1}^{i-1}) \tag{11}$$

其中, $W_{i-N+1}^{i-1} = W_{i-N+1}\cdots W_{i-1}$ 为词 W_i 的历史. 本文将利用基于词的二元语言模型和三元语言模型进行识别结果搜索, $N = 2$ 或 3. 概率 $P(\mathbb{C}|\mathbb{W})$ 可以表示为

$$P(\mathbb{C}|\mathbb{W}) = \prod_{k=1}^{l} P(C_k|W_k) \tag{12}$$

近似地认为所有的词都是等概率出现的, 对于同一个文本行图像, $P(C_k)$, $k = 1,\cdots,l$ 是一样的, 根据 Bayes 定理得到式 (13):

$$\hat{\mathbb{W}} = \arg\max_{W_1W_2\cdots W_l} \prod_{k=1}^{l} P(C_k|W_k)\prod_{k=1}^{l} P(W_k|W_{k-N+1}^{k-1}) \tag{13}$$

为了计算的方便, 取对数得式 (14):

$$\hat{\mathbb{W}} = \arg\max_{W_1W_2\cdots W_l} \left\{\sum_{k=1}^{l}\log P(C_k|W_k) + \sum_{k=1}^{l}\log P(W_k|W_{k-N+1}^{k-1})\right\} \tag{14}$$

组成词 W_k 的第 j 个字符为 w_k^j, w_k^j 为切分单元 c_k^j 的某个候选识别结果, 条件概率 $P(C_k|W_k)$ 为

$$P(C_k|W_k) = \prod_{j=1}^{m_k} P(c_k^j|w_k^j)$$

c_k^j 与字符类 w_k^j 的距离为 $d(c_k^j, w_k^j)$. 设字符特征服从高斯分布, 则 $d(c_k^j, w_k^j)$ 与条件概率 $P(c_k^j|w_k^j)$ 对数的负值成正比[12], 即

$$d(c_k^j, w_k^j) \propto -\log P(c_k^j|w_k^j)$$

将 $-\gamma\cdot d(c_k^j, w_k^j)$ 作为字符识别得分代入式 (14), 最终得到式 (15):

$$\hat{\mathbb{W}} = \arg\max_{W_1W_2\cdots W_l} \left\{-\gamma\sum_{k=1}^{l}\sum_{j=1}^{m_k} d(c_k^j, w_k^j) + \sum_{k=1}^{l}\log P(W_k|W_{k-N+1}^{k-1})\right\} \tag{15}$$

式 (15) 的最大化, 即候选识别结果的选择问题可以转化为在词图中搜索最优路径的问题. 切分单元的每个候选识别结果对应词图中的一个节点, 根据切分单元与候选字符类的距离得到节点字符识别得分. 基于词的二元及三元统计语言模型得到词图中节点之间的连接得分. 词图中最优路径上的词序列就是最终的字符识别结果. 为了构造词图, 首先对第 3 节处理得到的切分单元进行预处理, 然后根据切分单元的位置和候选识别结果构造词图, 最后在词图中搜索最优路径.

4.1 切分单元预处理

切分单元预处理的流程是: 首先进行切分单元的去重复处理, 然后合并切分单元的候选识别结果, 调整切分单元的位置, 最后添加一些虚拟切分单元, 为词图的构造做好准备.

步骤 1. 去重复处理. 队列 $RCcList$ 中的某些切分单元在 X 轴方向上的位置重合度高, 且距离最小的字符类相同, 去重复处理从中选出一个切分单元的候选识别结果用于构造词图, 而把其他切分单元从队列 $RCcList$ 中去除. 设第 i 个切分单元 $C(i)$ 的宽度为 $w(i)$, $lx(i)$ 为切分单元左边界在 X 轴方向上的坐标, 切分单元右边界在 X 轴方向上的坐标

为 $rx(i)$, $rx(i)=lx(i)+w(i)-1$. 定义切分单元在 X 轴方向上位置重合度的计算公式为

$$olp=\frac{\min(rx(i),rx(j))-\max(lx(i),lx(j))}{\max(cw(i),cw(j))} \quad (16)$$

遍历队列 $RCcList$ 中的两两切分单元, 根据式 (16) 计算切分单元在 X 轴方向上的位置重合度 olp, 如果 $olp>0.75$, 且与这两个切分单元距离最小的字符类相同, 则保留信度高的切分单元, 从队列 $RCcList$ 中去除信度低的切分单元.

步骤 2. 合并切分单元的候选识别结果. 遍历队列 $RCcList$ 中的两两切分单元, 根据式 (16) 计算切分单元的重合度 olp, 如果 $olp>0.75$, 则合并这两个切分单元的候选识别结果, 把两个切分单元的候选识别结果当成一个切分单元的候选识别结果, 合并后切分单元的位置为置信度更高的那个切分单元的位置.

步骤 3. 调整切分单元的位置. 设 $C(i)$ 和 $C(j)$ 为在 X 轴方向上连续排列的两个切分单元, 它们都是文本行图像的正确切分单元, 候选识别结果包含正确的字符, 且它们在 X 轴方向上位置的重合度 olp 不为 0, 但是一个很小的值. 如果不调整切分单元的位置, 则根据切分单元位置构造的词图中不存在从 $C(i)$ 的候选字符节点指向 $C(j)$ 的候选字符节点的有向边, 词图路径搜索将不可能得到正确的识别结果. 因此构造词图前, 遍历队列 $RCcList$ 中在 X 轴方向上连续排列的两两切分单元 $C(i)$ 和 $C(j)$, 如果它们在 X 轴方向上位置的重合度 $olp<0.15$, 则调整切分单元 $C(j)$ 左边界的位置, 令其左边界的位置为切分单元 $C(i)$ 右边界的位置加 1, 即 $lx(j)=rx(i)+1$.

步骤 4. 添加一些虚拟切分单元. 除最左边和最右边的切分单元, 遍历队列 $RCcList$ 中的每个切分单元. 设切分单元的左边界在 X 轴方向上的位置为 lx, 在切分单元的左边找离它最近的切分单元左边界, 其位置为 $preLx$. 如果在 $[preLx,lx)$ 的范围内找不到某个切分单元的右边界, 则增加一个虚拟切分单元. 虚拟切分单元左边界的位置为 $preLx$, 右边界的位置为 $lx-1$, 虚拟切分单元有一个空字符的候选识别结果, 且假设虚拟切分单元的字符识别得分为 $-\gamma\cdot\alpha\times(cx-preCx)/H$. 设切分单元的右边界在 X 轴方向上的坐标为 rx, 在切分单元的右边找离它最近的切分单元右边界, 其位置为 $aftRx$. 如果在 $(rx,aftRx]$ 的范围内找不到某个切分单元的左边界, 则增加一个虚拟切分单元. 虚拟切分单元左边界的位置为 $rx+1$, 右边界的位置为 $aftRx$, 虚拟切分单元有一个空字符的候选识别结果, 且假设虚拟切分单元的字符识别得分为 $-\gamma\cdot\alpha\times(aftEx-ex)/H$.

4.2　构造有向词图

每个切分单元的候选识别结果为图中的一个节点, 根据切分单元的候选识别结果构造有向图的步骤为:

步骤 1. 把切分单元的候选识别结果作为节点加入词图, 同时在图中添加一个起始节点和终止节点.

步骤 2. 设切分单元右边界共有 Q 个不同的位置, 则将词图分为 $P=Q+2$ 个单元, 右边界位置相同的切分单元及其候选识别结果节点在同一个单元; 起始节点和终止节点分别在第 0 单元和第 $P-1$ 单元.

步骤 3. 遍历每个切分单元, 对于第 j 个切分单元, 根据切分单元的位置在其左边界的左边寻找离其最近的切分单元, 最近切分单元所在单元为切分单元 j 的父单元. 如果在切分单元的左边没有切分单元, 则其父单元为 0. 图中每个候选识别结果节点的父单元为其对应切分单元的父单元, 起始节点没有父单元, 终止节点的父单元为 $P-2$.

步骤 4. 遍历图中每个节点, 对于第 k 个节点, 在节点 k 的父单元中的每个节点与节点 k 之间添加指向节点 k 的有向边.

步骤 5. 将图中相邻节点的字符组成词加入词图, 词的字符识别得分为组成词的字符的字符识别得分之和.

对图 6 中切分单元预处理后, 添加了一个虚拟切分单元, 则右边界共有 10 个不同的位置. 将词图分为 12 个单元, 起始节点和终止节点分别在第 0 单元和第 11 单元, 每个切分单元的候选识别结果节点在其对应的单元中, 根据节点之间的关系添加有向边, 最终构造的词图如图 7 所示. 图 7 中的粗边线矩形框为添加的虚拟切分单元节点.

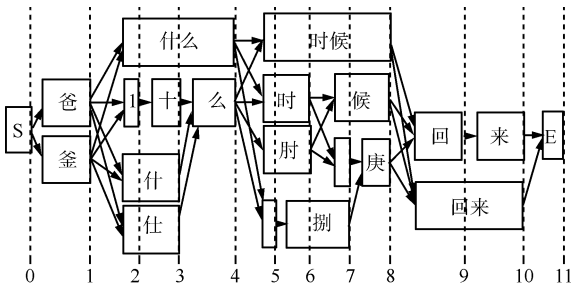


图 7　词图

Fig. 7　Word graph

4.3　词图最优路径搜索

基于词的二元及三元统计语言模型得到词图中节点之间的连接得分, 利用两阶段词图搜索算法寻找词图中的最优路径[13, 17]. 两阶段词图搜索首先基

于二元统计语言模型利用从前向后的 Viterbi-Beam 搜索算法进行第一阶段的词图搜索和剪枝, 然后基于三元统计语言模型利用从后向前的 A* 算法进行第二阶段的词图搜索, 找到 N-Best 路径. 把得分最高的路径上的词序列作为最终文本行图像的识别结果. 利用两阶段词图搜索算法寻找图 7 中的最优路径, 最优路径如图 8 所示, 路径上的词序列就是图像的正确识别结果.

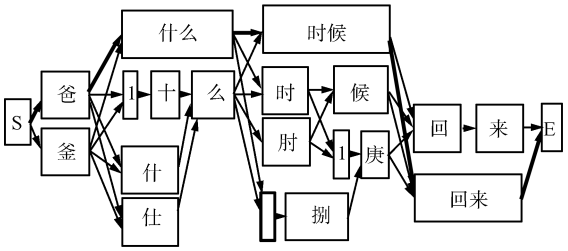


图 8 词图搜索得到的最优路径

Fig. 8 The best path found in the word graph

5 实验及结果分析

为了验证本文算法的有效性, 我们进行了相关实验. 首先将本文的集成型字符切分与识别算法与串行的字符切分与识别算法进行了对比实验, 然后分析了算法参数对字符识别率的影响.

对于单个切分单元的字符识别, 本文采用文献 [13] 中基于融合图像的字符识别方法. 该方法首先融合字符的二值图像和灰度得到融合图像, 然后对融合图像的大小和位置归一化, 提取 512 维的梯度直方图特征, 并经正则化的线性判别分析 (Linear discriminant analysis, LDA) 降维, 对降维后的特征归一化得到最终的特征, 分类器采用基于欧氏距离的最近邻法. 通过模拟视频中字符的退化过程为每种字体的每个字符生成了 30 个训练样本用于训练字符识别器. 利用基于新闻语料库训练好的统计语言模型计算词序列的概率[16].

我们对两个文本行图像集进行了测试: 电视文本行图像集和电影文本行图像集. 电视文本行图像集是从凤凰卫视的《凤凰早班车》, 中央电视台第一套的《新闻联播》, 中央电视台第一套的《新闻 30 分》, 浙江卫视的《浙江新闻联播》, 郑州卫视的《郑州新闻》, 遵义电视台的《遵义新闻联播》以及这些电视台的广告和天气预报中提取的文本行图像, 共计 1 724 幅文本行图像, 包含 13 422 个不同字体和大小的字符. 电影文本行图像集是从电影中提取的文本行图像, 共计 3 840 幅文本行图像, 包含 29 152 个字符.

5.1 算法性能比较

为了证明本文集成型算法的有效性, 与文献 [3] 和文献 [5] 中的串行算法进行了比较实验. 文献 [3] 中的字符切分是基于启发的方法; 文献 [5] 中的字符切分是基于字符识别器的方法. 实验中, 串行算法首先进行字符切分, 然后对切分单元进行字符识别, 最后基于统计语言模型从切分单元的候选字符类中搜索最终的识别结果.

在实验中, 本文集成型算法的参数为: $\alpha = 0.45$, $\beta = 1.25$, $\gamma = 40$. 串行算法每个切分单元的候选字符类数与本文算法的选取方法相同. 实验结果如表 1 所示. 从表 1 可以看出, 本文集成型算法的字符识别率明显高于串行的算法. 与文献 [5] 中的串行的算法相比, 对于第一个电视文本行图像测试集, 字符识别的错误率下降了 28.99 %, 对于第二个电影文本行图像测试集, 字符识别的错误率下降了 41.38 %.

通过分析发现, 本文算法的错误识别结果主要来源于下面两个因素: 1) 图像分辨率低, 文字相邻笔划粘连严重, 导致错误的连通域切分, 切分单元的候选识别结果不包含正确的字符; 2) 对于包含复杂背景的文本行图像, 其二值图像中很可能存在非字符的 “噪声” 成分, 且在处理中还可能与其他是字符的连通域合并, 导致切分单元的候选识别结果不包含正确的字符.

5.2 算法参数对字符识别性能的影响

影响本文集成型算法字符识别率的三个主要参数为: 1) 判断切分单元是否可能为字符的阈值 α; 2) 确定切分单元候选字符类数的参数 β; 3) 在后处理中平衡词之间的连接得分与字符识别得分的参数 γ. 在实验中, 我们选择了电视文本行图像集用于测试算法参数对字符识别性能的影响.

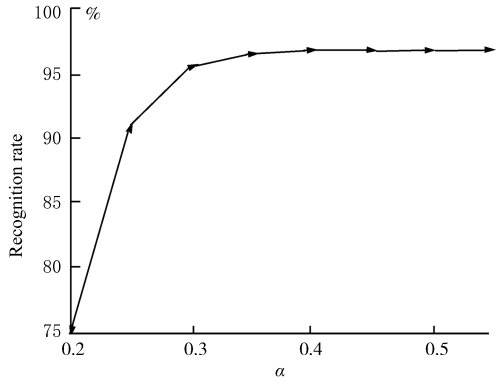
当参数 $\beta = 1.25$ 和 $\gamma = 40$ 保持不变时, 参数 α

表 1 视频文本行图像集字符识别性能

Table 1 Performance on text images extracted from videos

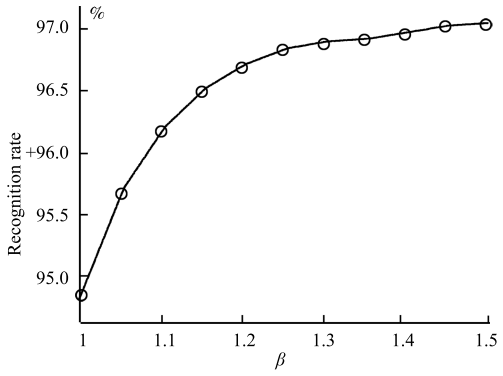
算法	电视文本行图像集		电影文本行图像集	
	字符识别率 (%)	处理时间 (s)	字符识别率 (%)	处理时间 (s)
文献 [3] 的串行算法	87.99	210.74	82.86	441.00
文献 [5] 的串行算法	95.55	440.46	98.26	918.79
本文算法	96.84	511.75	98.98	937.69

对字符识别率的影响如图 9 (a) 所示. 当 α 在 0.35 至 0.55 之间取值时, 字符识别率变化很小, α 约为 0.45 时, 字符识别率最高. 当 α 太小时, 可能把一些是字符的切分单元判断为非字符, 降低了字符识别率. 当 α 太大时, 可能把很多非字符的连通域或连通域的组合判断为字符, 则可能干扰基于语言模型的后处理, 影响字符识别率, 同时增大了计算量.



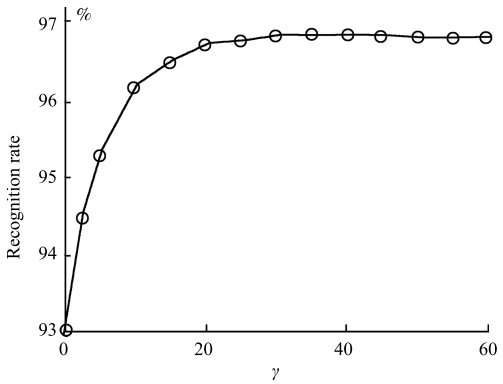
(a) 参数 α 对字符识别率的影响

(a) The influence of the parameter α on recognition rate



(b) 参数 β 对字符识别率的影响

(b) The influence of the parameter β on recognition rate



(c) 参数 γ 对字符识别率的影响

(c) The influence of the parameter γ on recognition rate

图 9 算法参数对字符识别率的影响

Fig. 9 The influences of the parameters on recognition rate

当参数 $\alpha = 0.45$ 和 $\gamma = 40$ 保持不变时, 参数 β 对字符识别率的影响如图 9 (b) 所示. 从图中可以看出, 当 β 为 1 时, 算法只选取与切分单元距离最小的字符类作为候选字符识别结果, 不能充分发挥语言模型候选识别结果选择的作用, 字符识别率最低. 随着参数 β 的增大, 切分单元的字符类以更大的概率出现在候选识别结果中, 更能发挥语言模型候选识别结果选择的作用, 提高了字符识别率.

当参数 $\alpha = 0.45$ 和 $\beta = 1.25$ 保持不变时, 参数 γ 对字符识别率的影响如图 9 (c) 所示. 从图中可以看出, 当 $\gamma = 0$ 时, 字符识别率最低, 当 γ 在 20 ∼ 55 之间取值时, 字符识别率变化很小, γ 约为 35 时, 词之间的连接得分与字符识别得分得到最好的平衡, 字符识别率最高.

6 结论

对于视频文本行图像的识别, 为了克服串行算法以及传统集成型算法的弱点, 本文提出了一种集成型的字符切分与识别处理算法. 该集成型算法能对粘连字符及复杂背景中的字符进行更准确的识别. 算法首先利用文本行图像的二值图估计文本行高度, 然后对二值图进行连通域分析, 对宽度较大的连通域进行切分处理得到宽度较小的连通域. 在连通域切分处理中, 根据字符笔划粘连的程度, 利用基于垂直投影的方法或基于滑动窗口的方法切分超宽连通域. 基于字符识别组合连通域, 最后基于统计语言模型搜索最优的识别结果. 实验结果表明, 与一种基于字符识别的串行算法相比, 本文集成型算法的字符识别错误率下降了 29 %, 证明了本文算法的有效性. 通过实验还分析了算法参数对字符识别率的影响.

References

1 Lienhart R, Wernicke A. Localizing and segmenting text in images and videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2002, **12**(4): 256−268

2 Lyu M R, Song J Q, Cai M. A comprehensive method for multilingual video text detection, localization, and extraction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2005, **15**(2): 243−255

3 Tang X O, Gao X B, Liu J Z, Zhang H J. A spatial-temporal approach for video caption detection and recognition. *IEEE Transactions on Neural Network*, 2002, **13**(4): 961−971

4 Liang S, Ahmadi M, Shridhar. Segmentation of touching characters in printed document recognition. In: Proceedings of the 2nd International Conference on Document Analysis Recognition. Tsukuba Science City, Japan: IEEE, 1993. 569−572

5 Yang W Y, Zhang S W, Zheng H B, Zeng Z. A recognition-based method for segmentation of Chinese character in images and videos. In: Proceedings of the International Conference on Audio, Language and Image Processing. Shanghai, China: IEEE, 2008. 723−728

6 Ha J, Haralick R M, Phillips I T. Recursive X-Y cut using bounding boxes of connected components. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Canada: IEEE, 1995. 952−955

7 Lee S W, Kim S Y. Integrated segmentation and recognition of handwritten numerals with cascade neural network. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 1999, **29**(2): 285−290

8 Casey R G, Nagy G. Recursive segmentation and classification of composite character patterns. In: Proceedings of the 6th International Conference on Pattern Recognition. Munich, Germany: IEEE, 1982. 1023−1026

9 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

10 Song J Q, Li Z, Lyu M R, Cai S J. Recognition of merged characters based on forepart prediction, necessity-sufficiency matching, and character-adaptive masking. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2005, **35**(1): 2−11

11 Wang Q F, Yin F, Liu C L. Integrating language model in handwriting Chinese text recognition. In: Proceedings of the 10th International Conference on Document Analysis and Recognition. Barcelona, Spain: IEEE, 2009. 1036−1040

12 Liu C L, Sako H, Fujisawa H. Effects of classifier structures and training regimes on integrated segmentation and recognition of handwritten numeral strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(11): 1395−1407

13 Yang Wu-Yi. Text Extraction in Video [Ph. D. dissertation], Institute of Automation, Chinese Academy of Sciences, China, 2009
(杨武夷. 视频文字信息抽取技术研究 [博士学位论文]. 中国科学院自动化研究所, 中国, 2009)

14 Tsujimoto S, Asad H. Major components of a complete text reading system. *Proceedings of the IEEE*, 1992, **80**(7): 1133−1149

15 Liu C L, Nakagawa M. Precise candidate selection for large character set recognition by confidence evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, **22**(6): 636−641

16 Wang Xiao-Rui. Data Clustering Based Language Modeling for Speech Recognition [Ph. D. dissertation], Institute of Automation, Chinese Academy of Sciences, China, 2008
(王晓瑞. 基于数据聚类的语言建模研究 [博士学位论文]. 中国科学院自动化研究所, 中国, 2008)

17 Soong F K, Huang E F. A tree-trellis based fast search for finding the N-best sentence hypotheses in continuous speech recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Toronto, Canada: IEEE, 1991. 705−708

杨武夷 厦门大学海洋与环境学院助理教授. 2009 年获中国科学院自动化研究所工学博士学位. 主要研究方向为信号处理、图像处理、字符识别和模式识别. 本文通信作者.
E-mail: eagleywy@126.com
(**YANG Wu-Yi** Assistant professor at the College of Oceanography and Environmental Science, Xiamen University. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 2009. His research interest covers signal processing, image processing, character recognition, and pattern recognition. Corresponding author of this paper.)

张树武 中国科学院自动化研究所研究员. 1997 年获中国科学院自动化研究所博士学位. 主要研究方向为数字媒体技术、多语言语音识别和自然语言处理.
E-mail: swzhang@hitic.ia.ac.cn
(**ZHANG Shu-Wu** Professor at the Institute of Automation, Chinese Academy of Sciences. He received his Ph. D. degree from the Institute of Automation, Chinese Academy of Sciences in 1997. His research interest covers digital media technologies, multilingual speech recognition, and natural language processing.)