软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(12):2919−2933 [doi: 10.3724/SP.J.1001.2011.03970]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于颜色聚类和多帧融合的视频文字识别方法*

易 剑, 彭宇新+, 肖建国

(北京大学 计算机科学技术研究所,北京 100871)

Recognition of Text in Video Based on Color Clustering and Multiple Frame Integration

YI Jian, PENG Yu-Xin+, XIAO Jian-Guo

(Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

+ Corresponding author: E-mail: pengyuxin@pku.edu.cn

Yi J, Peng YX, Xiao JG. Recognition of text in video based on color clustering and multiple frame integration. *Journal of Software*, 2011,22(12):2919–2933. http://www.jos.org.cn/1000-9825/3970.htm

Abstract: This paper proposes a new approach for the text recognition of video, whose novelty mainly lies in the color-based clustering and multiple frame integration of three phases: First, in the text detection phase, the two significant features of text block are jointly considered in a video: homogeneous color, dense edges, and color-based clustering are employed to decompose the color edge map of video frame into several edge maps, which make the text detection more accurate. Second, in text enhancement phase, the text blocks are identified and integrated with the same content by filtering the blurred text based on the proposed text-intensity map, which can obtain the clean background and clear text with a high contrast of effective text extraction. Third, in the text extraction phase, on one hand, for effective binarization of text block, instead of performing binarization in a constant color plane as in the existing methods, this approach can adaptively select the best color plane according to the text contrast difference among color planes for binarization. On the other hand, for effective text recognition, the color differences between the text and background in video frames are considred, and color-based clustering is utilized to remove the noises. Extensive experimental results have shown that this approach outperforms several existing state-of-the- art methods.

Key words: text recognition of video; color-based clustering; multiple frame integration; video retrieval; noise removal

摘 要: 提出一种基于颜色聚类和多帧融合的视频文字识别方法,首先,在视频文字检测模块,综合考虑了文字区域的两个显著特征:一致的颜色和密集的边缘,利用近邻传播聚类算法,根据图像中边缘颜色的复杂程度,自适应地把彩色边缘分解到若干边缘子图中去,使得在各个子图中检测文字区域更为准确.其次,在视频文字增强模块,基于文字笔画强度图过滤掉模糊的文字区域,并综合平均融合和最小值融合的优点,对在不同视频帧中检测到的、包含相同内容的文字区域进行融合,能够得到背景更为平滑、笔画更为清晰的文字区域图像.最后,在视频文字提取模块,通过自适应地选取具有较高文字对比度的颜色分量进行二值化,能够取得比现有方法更好的二值化结果;另一方面,基于图像中背景与文字的颜色差异,利用颜色聚类的方法去除噪声,能够有效地提高文字识别率.实验结果表明,该方法能够比现有方法取得更好的文字识别结果.

* 基金项目: 国家自然科学基金(60873154, 61073084); 国家发改委资助项目([2010]3044)
收稿时间: 2009-11-25; 修改时间: 2010-09-08; 定稿时间: 2010-11-03

关键词: 视频文字识别;基于颜色的聚类;多帧融合;视频检索;噪声去除
中图法分类号: TP391 文献标识码: A

随着互联网技术与多媒体技术的迅速发展,网络上出现了海量的视频内容.如何对海量的视频信息建立索引,使用户能够迅速检索到想要的内容,成为了一个亟待解决的关键问题.传统的方法基于人工标注的关键词进行检索,但不能适用于海量视频内容的分析与检索.另一方面,大量的视频内容中含有文字信息,这些文字信息一般同视频的内容密切相关,能够对其进行描述.因此,如果能够准确识别视频中的文字,将会极大地促进视频内容分析与检索技术的发展.一般而言,现有的视频文字识别方法主要包含 4 个模块:视频文字检测模块、视频文字增强模块、视频文字提取模块和 OCR 软件识别模块.其中:视频文字检测模块对视频内容进行了分析,在视频帧中检测和定位文字区域;视频文字增强模块主要采用多帧融合方法,对在多个视频帧中检测到的相同文字区域进行融合,以得到背景更为平滑、笔画更为清晰的文字图像;视频文字提取模块对文字区域图像进行处理,把文字从背景中分割出来,转化成可供 OCR 软件识别的二值文字图像;OCR 软件识别模块识别二值文字图像,完成图像文字到文本的转换.在这 4 个模块中,OCR 是比较成熟的技术,在市场上已有成功的应用.因此,现有研究主要集中在视频文字检测、基于多帧融合的视频文字增强和视频文字提取这 3 个模块上.

在视频文字检测模块,文字区域一般具有两个明显的视觉特征:密集的文字边缘和一致的文字颜色.现有方法[1−4]大多采用了这两个特征之一.例如,文献[1]利用文字区域的边缘特征检测文字,该方法对视频帧的边缘图进行形态学处理,使边缘图中的密集边缘形成多个初始连通分量,然后对连通分量进行分析以确定文字区域.与基于边缘特征的方法类似,基于颜色特征的方法[2]也采用了连通分量分析.不同之处在于,基于颜色特征的方法是通过对图像中的颜色进行分析来产生初始连通分量.此外,还有一些方法[3,4]把视频中的文字看作是一种特殊的纹理,利用机器学习的方法把图像中的像素点分为文字像素和背景像素.这些方法的不足之处在于没有考虑把边缘和颜色特征结合起来.在视频文字增强模块,现有研究大多采用多帧融合来增强文字笔画的清晰程度[5−9].有的采用了平均融合[5,6],这些方法能够平滑复杂的背景,但不能提高文字与背景的对比度;另一些方法则采用了最小值融合[7−9],这些方法可以提高文字与背景的对比度,但容易受到噪声的影响.文献[6]中的方法只选取具有较高对比度的文字区域进行融合,然而,由于受到网络带宽的限制,许多视频尤其是网络视频被压缩到了较小的码率,使得视频图像的质量受到了很大的影响,一些文字区域的笔画变得模糊不清,导致了较为模糊的文字融合结果.如果能够识别和舍弃模糊的文字区域,并综合平均融合与最小值融合的优点,将非常有利于得到清晰的文字笔画融合结果.在视频文字提取模块,现有方法主要解决两个关键问题[10−12]:二值化和噪声去除.对于二值化问题,最常用的方法是基于阈值的方法[10],首先在彩色文字区域图像的某一个固定颜色分量上取得灰度图像,然后对灰度图像统计全局阈值或局部阈值进行二值化.然而,这种方法并不一定合理,对于不同颜色的文字,在不同的颜色分量上进行二值化可能会取得更好的结果.对于噪声去除问题,有的方法考虑了连通分量的形状特征[10−11],去除那些形状上与文字笔画差异较大的噪声;有的方法则没有考虑这个问题[12],当噪声形状与文字类似时,就不能被去除,极大地影响了文字识别的结果.这些方法都没有考虑到图像文字与背景之间的颜色差异,利用这个差异,可以有效去除噪声.基于上述分析和考虑,针对现有方法在各个模块的不足,本文提出了一种基于颜色聚类和多帧融合的视频文字识别方法,主要的创新之处是:(1) 在视频文字检测模块,提出了彩色边缘分层的思想,把文字的密集边缘和一致颜色两个特征结合起来,根据图像中边缘颜色的复杂程度,用近邻传播聚类算法自适应地把彩色边缘分解到多个边缘子图中,使得在各个子图中进行文字检测更为简单和准确;(2) 在视频文字增强模块,我们用文字笔画强度图描述图像中文字笔画的清晰程度,过滤掉模糊的文字区域,并综合使用平均融合与最小值融合,能够取得更好的多帧融合结果;(3) 在视频文字提取模块,一方面根据彩色文字区域图像在不同颜色分量上文字边缘的强弱,选择具有较强对比度的颜色分量进行二值化,能够取得比现有二值化方法[10]更好的结果;另一方面,本文考虑了文字与背景之间的颜色差异,采用基于颜色聚类的方法,并结合连通分量分析[10]和灰度一致性分析[11]去除噪声,能够有效提高文字识别率.

1 总体框架

如图 1 所示,本文方法主要由 4 个模块组成:视频文字检测模块(video text detection)、视频文字增强模块(video text enhancement)、视频文字提取模块(video text extraction)和 OCR 软件识别模块(OCR recognition).在这 4 个模块当中,OCR 软件识别模块采用了现有方法识别文字,因此,本文的工作主要集中在其他 3 个模块上.下面将对我们的方法进行详细描述.

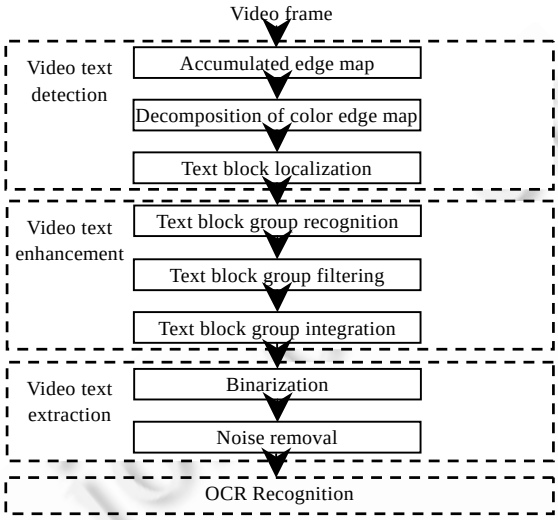


Fig.1 Framework of our method for text recognition of video

图 1 本文的视频文字识别方法总体框架

2 视频文字检测

本文采用近邻传播聚类算法[13]把原图中的边缘分解到若干个子图中,这样,具有不同颜色的文字边缘和背景边缘被分开,不同颜色的文字边缘也被分开,在各个子图中检测和定位文字区域也相对准确和容易.与传统的聚类方法相比,近邻传播聚类算法有两个优点:(1) 传统的聚类算法如 *K*-means 等需要预先指定 *k* 的值,然而,视频中的颜色是不可预知的.近邻传播聚类算法可以根据视频中颜色的复杂程度自动确定分类的类数,更为合理;(2) *K*-means 算法的聚类效果和初始种子点的选取密切相关,而近邻传播聚类算法把参与聚类的结点都看作是聚类中心的候选结点,通过在结点之间传播信息找到最终的聚类中心,结果更为稳定.本文的视频文字检测方法主要包含 3 个步骤:累积边缘图生成(accumulated edge map generation)、彩色边缘分层(decomposition of color edge map)和文字区域定位(text block localization).下面详细描述各个步骤的细节.

2.1 累积边缘图生成

现有的大多数边缘检测方法在彩色图像的 *Y* 分量上进行边缘检测[10],然而这种方法并不一定合理,有些在彩色图像上很明显的文字边缘,在 *Y* 分量上却非常模糊.因此,本文定义了累积边缘图.累积边缘图由在原图的 *YUV* 各个分量上检测到的边缘图合并得到.同在单个颜色分量上检测到的边缘图相比,累积边缘图中的文字边缘信息更为丰富,也更利于文字的检测.设原图为 *I*,*I* 的累积边缘图 *E* 由公式(1)表示:

$$E(x,y)=\min(E_Y(x,y)+E_U(x,y)+E_V(x,y),255) \tag{1}$$

其中,E_Y,E_U 和 E_V 是在图像的 *YUV* 分量上分别检测到的边缘图,它们由公式(2)计算得到:

$$E_\alpha=\max(S_H,S_V,S_{LD},S_{RD}),\alpha\in\{Y,U,V\} \tag{2}$$

在公式(2)中,S_H,S_V,S_{LD} 和 S_{RD} 分别是用 *Sobel* 边缘检测算子得到的水平、垂直、左对角线和右对角线的边

缘强度值.在图像中,文字边缘一般较为明显,具有较大的强度值.因此,在本文中,若边缘图中点的强度值小于 T_{back},则认为是背景边缘,并把相应的 $E(x,y)$置为 0.

2.2 彩色边缘分层

为了在边缘图中准确地定位文字区域,我们考虑了文字边缘和背景边缘的颜色差异,并利用文献[13]中的近邻传播聚类算法把累积边缘图中的边缘分解到若干较为简单的边缘子图中.具体做法是:我们对上一节中得到的累积边缘图 E 进行着色,着色的方法见公式(3).其中,E'是 E 对应的彩色边缘图,E'中边缘点的颜色值 $E'_{RGB}(x,y)$ 被置为该边缘点在原图 I 中相应的颜色 $I_{RGB}(x,y)$,E'中的非边缘点被置为黑色.

$$E'_{RGB}(x,y)=\begin{cases} I_{RGB}(x,y), & E(x,y)>0 \\ (0\ 0\ 0), & E(x,y)=0 \end{cases} \tag{3}$$

然后采用文献[13]中提出的近邻传播聚类算法,把 E'中的点根据其颜色值的不同分解到若干边缘子图中.近邻传播聚类算法定义了两类信息量,这两类信息量分别用代表矩阵 $R=[r(i,k)]_{n\times n}$和适选矩阵 $A=[a(i,k)]_{n\times n}$表示.其中:$r(i,k)$表示结点 x_k 作为结点 x_i 的类代表点的能力;$a(i,k)$表示选择结点 x_i 作为结点 x_k 的类代表点的合适程度.算法按公式(4)~公式(6)不断迭代,更新这两类信息的值得到聚类结果,算法的细节见文献[13].

$$r(i,k)\leftarrow s(i,k)-\max_{k'\ s.t.\ k'\neq k}\{a(i,k')+s(i,k')\} \tag{4}$$

$$\text{if } i\neq k,\ a(i,k)\leftarrow \min\{0,r(k,k)+\sum\nolimits_{i'\ s.t.\ i'\{i,k\}}\max(0,r(i',k))\} \tag{5}$$

$$a(k,k)\leftarrow \sum\nolimits_{i'\ s.t.\ i'\neq k}\max\{0,r(i',k)\} \tag{6}$$

假设 E'中的边缘点包含 C 种颜色,每种颜色对应近邻传播聚类算法中的一个结点,n 的取值就等于 C,即颜色种类的数量.$s(i,k)$定义为 $s(i,k)=-dis(i,k),i\neq k$.其中,$dis(i,k)$是颜色 i 与 k 的差异值,这里定义为颜色 i 与 k 对应的由 RGB 这 3 个颜色分量组成的向量之间的欧式距离.$s(i,k)$值越大,表示颜色 i 与 k 越相近;$s(i,k)$值越小,表示颜色 i 与 k 差异越大.$s(k,k)$的值赋为不同颜色 $s(i,k)$值中的最小值.聚类的结果是,C 种颜色被分为 l 类,颜色属于同一类的像素点被分解到同一边缘子图中.l 的值由近邻传播聚类算法根据边缘像素点的颜色复杂度自动确定.

2.3 文字区域定位

一般情况下,文字区域的边缘比较密集,通过对边缘子图进行水平和垂直投影可以定位文字区域.因此,本文采取了类似于文献[10]中的投影方法在各个边缘子图中定位文字区域.然而,这样检测到的文字区域中还包含一些错误的文字区域,为了去除这些错误的文字区域,本文采用了文献[11]中基于 SVM 的过滤方法,把定位到的文字区域分成正确和错误的文字区域,并舍弃错误的文字区域,这样能够极大地提高文字区域检测的准确性.本文测试了多种特征,根据实验结果,采取的特征简单描述为:对图像检测得到 0°方向、45°方向、90°方向以及 135°方向上的边缘图,这样就得到了 4 个方向的边缘图.分别在这些边缘图上计算均值、方差、能量、熵、惯量以及局部同次性这 6 个特征,4 个边缘图就可以得到 24 维的特征.

3 视频文字增强

本文提出一种基于文字笔画强度图的多帧融合方法,主要包含 3 个步骤:文字区域组识别(text block group recognition)、文字区域组过滤(text block group filtering)和文字区域组融合(text block group integration).与现有方法相比,本文方法定义了文字笔画强度图,用于衡量图像中文字笔画的清晰程度,只选取那些含有清晰笔画的文字区域进行融合,避免了模糊文字区域对于融合结果的影响.下面详细说明本文方法的各个步骤.

3.1 文字区域组识别

为识别含有相同内容的多个文字区域,除了使用文献[6]中的位置特征,本文还考虑了其他两种特征,即文字区域的边缘分布特征和对比度特征.不同的文字区域只有在这 3 个特征上都相似时,才被认为包含了相同的文字.这样做的原因包括如下两点:(1) 文字区域中的边缘主要是文字边缘,如果两个区域具有相同的文字,那么它们应该具有类似的边缘分布;(2) 文字区域中的对比度主要是文字与背景的颜色差异,因此含有相同文字的区

域应该具有类似的对比度.图 2 给出了本文方法的细节.

1. $SimilarLoc=SimilarEdgeDis=SimilarCon=$False
2. If $Overlap(t_a,t_b)>r_1\times\mathrm{Min}(area(t_a),area(t_b))$
   $SimilarLoc=$True
3. $NoneZero(E_a,E_b)=\{p|E_a(p)>0\ \&\ E_b(p)>0\}$
   If $NoneZero(E_a,E_b)>r_2\times Overlap(t_a,t_b)$
   $SimilarEdgeDis=$True
4. $EdgeIDiff(t_a,t_b)=Sum(|E_a(p)-E_b(p)|)$
   If $EdgeIDiff(t_a,t_b)<D_{MAX}\times Overlap(t_a,t_b)$
   $SimilarCon=$True
5. If $SimilarLoc$ & $SimilarEdgeDis$ & $SimilarCon$
   t_a and t_b contain the same text
   Else
   t_a and t_b contain the same text

Fig.2 Our method for text block group recognition

图 2 本文的文字区域组识别算法

图 2 的 t_a 和 t_b 表示在连续视频帧中检测到的两个文字区域.在第 2 步中,$Overlap(t_a,t_b)$是区域 t_a 与 t_b 的重叠部分,r_1 是取值范围为 0~1 的一个常数,只有在 $Overlap(t_a,t_b)$足够大的时候,*SimilarLoc* 的值被置为 True,表示区域 t_a 与 t_b 在不同视频帧中处于相近的位置.在第 3 步中,E_a 与 E_b 是 t_a 和 t_b 的边缘图,p 是 $Overlap(t_a,t_b)$中的像素,r_2 是取值范围 0~1 之间的一个常数,$NoneZero(E_a,E_b)$是在 E_a 与 E_b 中边缘强度值都不为 0 的像素集合,能够用来衡量 E_a 与 E_b 中的边缘分布情况.当 $NoneZero(E_a,E_b)$中包含的像素个数大于 r_2 与 $Overlap(t_a,t_b)$的乘积时,*SimilarEdgeDis* 被设为 True,表示 E_a 与 E_b 具有类似的边缘分布.在第 4 步中,D_{MAX} 是一个阈值,$EdgeDiff(t_a,t_b)$是边缘图 E_a 与 E_b 中相应边缘点强度值之差的累加和.因为边缘图 E_a 与 E_b 能够描述 t_a 和 t_b 之中的对比度情况,因此 $EdgeDiff(t_a,t_b)$可以用来描述 t_a 和 t_b 之间的对比度差异.只有在 $EdgeDiff(t_a,t_b)$的值小于 D_{MAX} 与 $Overlap(t_a,t_b)$的乘积时,*SimilarCon* 才被置为 True,即 t_a 与 t_b 有相似的对比度.第 5 步说明,只有在 *SimilarLoc*,*SimilarEdgeDis* 和 *SimilarCon* 的值都为 True,即 t_a 与 t_b 具有 3 个相似特征的时候,才被认为含有相同的文字内容.图 3 给出了一个例子,其中,图 3(a)和图 3(b)是在两个连续视频帧中检测到的两个文字区域,这两个文字区域处于视频帧中的相同位置,并且含有相同的字数和相近的宽度,按照文献[6]中的方法,它们被识别为相同的文字区域,显然,这样会得到错误的融合结果.图 3(c)和图 3(d)是在这两个视频帧中检测到的两个文字区域.另一方面,如图 3(e)和图 3(f)所示,不同的文字内容具有不同的边缘分布和强度,通过比较文字区域的边缘图,可以区别出不同的文字内容.因此在本文中,我们不仅考虑了文字区域的位置和形状信息,同时通过比较文字区域内的边缘分布和强度值,能够更为有效地区分出不同内容的文字区域.

(a) (b)

Fig.3

图 3

3.2 文字区域组过滤

由于网络带宽的限制,许多视频,尤其是网络视频经过较大的压缩,码率较低,影响了图像的质量;而叠加的文字在视频中是以图像形式存在的,这使得有些文字笔画十分模糊,这些文字会使多帧融合的结果模糊不清.图

4 给出了一些模糊文字区域的例子.

Fig.4 Several examples for blurred text in video

图 4 模糊文字区域例子

为了避免这些模糊文字带来的影响,本文定义了文字笔画强度图,用来描述视频帧中文字笔画的清晰程度,只选取含有较为清晰文字笔画的区域进行融合.文字笔画强度用公式(7)定义,该公式表示图像中位置(x,y)的文字笔画强度,其中,$TInt_i^H$,$TInt_i^V$,$TInt_i^{LD}$ 和 $TInt_i^{RD}$ 分别是用 4 个文字笔画检测算子在水平、垂直、左对角线和右对角线上计算得到强度值.图 5 展示了这 4 个检测算子(左边是示意图,右边是实验中采用的模板),清晰的文字笔画在文字笔画强度图中具有较高的强度值,而模糊的文字笔画具有较低的强度值,可以利用文字笔画在文字笔画强度图中的强弱程度表示其在原图中的清晰程度.强度值越大,笔画的清晰度就越高.

$$TIMap_i(x,y)=\mathrm{Max}(TInt_i^H,TInt_i^V,TInt_i^{LD},TInt_i^{RD}) \tag{7}$$

(a) Horizontal text-intensity detector
(a) 水平方向的文字笔画强度检测算子

(b) Vertical text-intensity detector
(b) 垂直方向的文字笔画强度检测算子

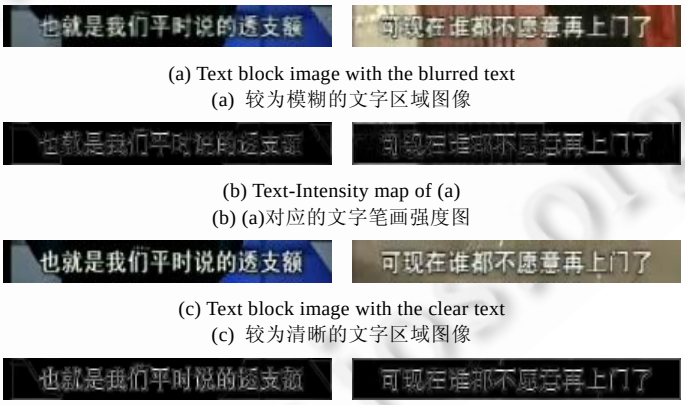
(c) Left diagonal text-intensity detector
(c) 左对角线方向上的文字笔画强度检测算子

(d) Right diagonal text-intensity detector
(d) 右对角线方向上的文字笔画强度检测算子

Fig.5

图 5

图 6 展示了两幅相同的文字区域图像,具有不同清晰程度的文字笔画.清晰的文字笔画在文字笔画强度图中具有较大的强度值,而模糊的文字笔画在文字笔画强度图中的强度值较低.



(a) Text block image with the blurred text
(a) 较为模糊的文字区域图像

(b) Text-Intensity map of (a)
(b) (a)对应的文字笔画强度图

(c) Text block image with the clear text
(c) 较为清晰的文字区域图像

(d) Text-Intensity map of (c)
(d) (c)对应的文字笔画强度图

Fig.6

图 6

本文的文字区域组过滤方法描述为:首先计算得到文字区域组内任一文字区域 t_i 的文字笔画强度图 $TIMap_i$,然后把 $TIMap_i$ 分成两个部分:文字部分 $TIMap_i^{text}$ 和背景部分 $TIMap_i^{back}$,并在文字部分 $TIMap_i^{text}$ 上计算 t_i 的文字笔画清晰程度 $TextClarity_i$.$TextClarity_i$ 越高,表示文字越清晰.反之,文字越模糊.过滤时去除那些 $TextClarity_i$ 值较小的模糊文字区域,这样,假设过滤之前文字区域组中包含 M 个文字区域,记为 $t_1,t_2,\ldots,t_M$;过滤

后,文字区域组剩下 M'个 $TextClarity_i$ 值较大的文字区域,记为 $t'_1,t'_2,...,t'_{M'}$.图 7 给出了我们的算法.在第 1 步中,$TInt_i^H$,$TInt_i^V$,$TInt_i^{LD}$ 和 $TInt_i^{RD}$ 分别表示用图 6 中的 4 个文字笔画强度检测算子在原图上检测到的水平、垂直以及两个对角线方向的文字笔画强度.在第 2 步中,t_{Avg} 是对 $t_1,t_2,\ldots,t_M$ 进行平均得到的图像,H_{otsu} 是在 t_{Avg} 中用 OTSU 方法求得的局部阈值,p 是 $TIMap_i$ 中的一个像素,$t_{Avg}(p)$是这个像素在其于 t_{Avg} 中相应位置的强度值.如果 $t_{Avg}(p)$大于 H_{otsu},则认为该 p 属于文字;否则属于背景.这是因为在视频中,相对于背景,文字一般具有较高的强度值.在第 3 步中,$TextClarity_i$ 表示图像中的文字笔画清晰程度,$TextClarity_i$ 值越大,表示 t_i 中的笔画越清晰.最后,我们在第 4 步中选取具有较高文字清晰度的 M'个文字区域图像进行融合.

1. For each $t_i(1\leqslant i\leqslant M)$,
   $TIMap_i=\mathrm{Max}(TInt_i^H,TInt_i^V,TInt_i^{LD},TInt_i^{RD})$
2. $t_{AVG}=AVG(t_1,t_2,\ldots,t_M)$
   For each $TIMap_i$
   $TIMap_i^{text}=\{p|p\in TIMap_i,t_{AVG}(p)>H_{otsu}\}$
   $TIMap_i^{back}=\{p|p\in TIMap_i,t_{AVG}(p)\leqslant H_{otsu}\}$
3. For each $t_i(1\leqslant i\leqslant M)$,
   $TextClarity_i=\sum_{p\in TIMap_i^{text}} TIMap_i(p)\Big/|TIMap_i^{text}|$
4. Select the M' blocks with highest text clarity for the integration

Fig.7 Our method for text block group filtering

图 7 本文的文字区域组过滤算法

3.3 文字区域组融合

本文首先取局部阈值把文字区域中的像素粗略地分为文字部分和背景部分.然后,在文字部分进行平均融合,这样不易受到噪声的影响,能够得到清晰的文字笔画;另一方面,在背景部分采取了最小值融合,这样能够达到简化背景和提高对比度的目的,更利于进行文字提取及识别.用 $t'_1,t'_2,...,t'_{M'}$ 表示从 $t_1,t_2,\ldots,t_M$ 中挑选出来的、含有较为清晰的文字笔画的文字区域,这种融合方法可以用公式(8)和公式(9)来描述:

$$t_{\mathrm{int}}(p)=\begin{cases}\min\{t'_i(p)\},\ 1\leqslant i\leqslant M',\ p\in t_i'^{back}\\ \sum_{1\leqslant i\leqslant M'} t'_i(p)/M',\qquad p\in t_i'^{text}\end{cases}\tag{8}$$

$$t_i'^{text}=\{p\mid p\in t'_i,t_{Avg}(p)>H_{otsu}\},t_i'^{back}=\{p\mid p\in t'_i,t_{Avg}(p)\leqslant H_{otsu}\}\tag{9}$$

其中,t_{int} 是多帧融合的结果,$t_i'^{text}$ 和 $t_i'^{back}$ 分别代表 t'_i 的文字部分和背景部分,$1\leqslant i\leqslant M'$,它们由公式(8)计算得到.公式(8)中:t_{Avg} 是对文字区域 $t_1,t_2,\ldots,t_M$ 进行平均得到的图像;H_{otsu} 是在 t_{Avg} 中用 OTSU 方法[14]求得的局部阈值;p 表示文字区域 t'_i 的一个像素,$t_{Avg}(p)$是这个像素在 t_{Avg} 中相应位置的强度值.如果 $t_{Avg}(p)>H_{otsu}$,则认为该 p 属于文字部分 $t_i'^{text}$;否则,属于背景部分 $t_i'^{back}$.

4 视频文字提取

视频文字提取模块对文字区域图像进行处理,把文字从复杂的图像背景中分割出来,得到可供 OCR 软件识别的二值文字图像.基于对现有方法不足的考虑,一方面,本文根据文字区域图片在 *YUV* 各个分量上对比度的不同,选取在对比度最大的颜色分量上进行二值化(binarization),能够取得比现有方法更好的结果;另一方面,本文考虑了文字笔画与噪声之间的颜色差异,利用基于颜色聚类的方法,并结合连通分量分析[10]以及灰度一致性分析[11]来去除噪声(noise removal),能够比现有方法取得更好的噪声去除结果.

4.1 基于颜色分量选择的二值化

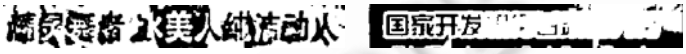
现有方法大多选择在彩色图像的 Y 分量上进行二值化[10],然而这并不总是最好的选择.随着图像文字颜色的变化,有时候在其他颜色分量上能够比在 Y 分量上取得更好的结果.图 8(a)是一个文字区域;图 8(b)和图 8(c)

分别是这个文字区域的 Y 和 U 分量;图 8(d)和图 8(e)分别是这个文字区域在 Y 和 U 分量上用文献[14]中的二值化方法处理得到的二值图片.显然,U 分量上的文字区域比 Y 分量上的对比度更高,背景也更为平滑,在 U 分量上的二值化结果好于在 Y 分量上二值化的结果.

(a) Text block
(a) 文字区域

(b) Y color component of (a)
(b) 图(a)的 Y 分量

(c) U color component of (a)
(c) 图(a)的 U 分量



(d) Binarized result of (b)
(d) 对图(b)进行二值化的结果

(e) Binarized result of (c)
(e) 对图(c)进行二值化的结果

Fig.8
图 8

进一步分析,在图 8(a)中,文字是红色的,即文字在 R 分量上的值较大,而在 G 和 B 分量上的值较小.另一方面,如果背景是其他颜色,即背景区域在 R 分量上的值比较小,而在 G 和 B 分量上的值较大.当把这张图片从 RGB 颜色空间转换到 YUV 颜色空间的时候,根据公式(10),V 分量上的值主要由 R 分量上的值决定,所以红色的文字在 V 分量上有较大的值,而其他颜色的背景区域在 V 分量上有较小的值.因此对于这张图片,在 V 分量上有两个优点:一是文字的对比度较大,二是背景简单.在 V 分量上对这张图片进行二值化更合适.

$$V=0.615\times R-0.515\times G-0.100\times B \tag{10}$$

基于上述分析,我们对文字区域图像在 YUV 空间各个颜色分量上的对比度进行比较,选取具有最高对比度的颜色分量进行二值化.公式(11)中,C_Y,C_U 和 C_V 代表了文字区域分别在 YUV 空间各个颜色分量上的对比度,C_α 是它们中最大的值,$\alpha\in\{Y,U,V\}$.因此我们选择在 YUV 空间中的 α 分量上进行二值化.公式(12)中,E_Y,E_U 和 E_V 是文字区域在 YUV 空间各个颜色分量上的边缘强度图.因为文字总是出现在文字区域的中央部分,所以我们将处于边缘强度图 E_α 中央部分的边缘强度值进行累加得到 $C_\alpha,\alpha\in\{Y,U,V\}$.因为边缘强度的高低代表原图对比度的大小,这样,$C_\alpha$ 就能代表文字区域在 α 分量上的对比度.C_α 值最大,就代表文字区域在 α 分量上的对比度最大,应该选择在 α 分量上进行二值化.本文采用了文献[15]中提出的经过改进的 Niblack 方法来对图像进行二值化.对于图像中的每一个像素,该方法根据该像素周围区域的灰度变化情况自适应地计算得到一个局部阈值,根据这个阈值来判断每一个像素是否属于文字.

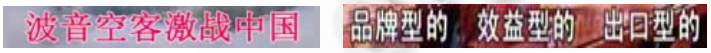
$$C_\alpha=\max(C_Y,C_U,C_V),\alpha\in\{Y,U,V\} \tag{11}$$

$$C_Y=\sum_{\substack{w/3\leqslant x\leqslant w\times 2/3\\ h/3\leqslant y\leqslant h\times 2/3}}E_Y(i,j),\ C_U=\sum_{\substack{w/3\leqslant x\leqslant w\times 2/3\\ h/3\leqslant y\leqslant h\times 2/3}}E_U(i,j),\ C_V=\sum_{\substack{w/3\leqslant x\leqslant w\times 2/3\\ h/3\leqslant y\leqslant h\times 2/3}}E_V(i,j) \tag{12}$$

4.2 基于颜色聚类的噪声去除

经过前面的处理,文字区域被二值化成了两部分:前景和背景.如图 9(b)所示,前景是一个连通分量的集合,

包括了文字的笔画和噪声.为了能够有效地识别文字,有必要在二值化之后进行噪声去除的处理.现有的噪声去除方法包括连通分量分析[10]和灰度一致性分析[11].然而这两种方法并不总是有效的,例如,当噪声块的几何形状和灰度值都与文字笔画很接近的时候,这样的噪声就不能被除去,如图 9(c)所示.一般来说,在图片中,文字的笔画总是有很相近的颜色,而噪声块在图片中的颜色和文字笔画总是不同的,如图 9(d)所示.因此,基于文字笔画与噪声的颜色差异,我们用聚类的方法把连通分量分为两类:一类是文字笔画,另一类是噪声.这样,噪声的那一类就被除去,而笔画的那一类则被保留,如图 9(e)所示.基于上述分析,本文把基于颜色聚类的方法和连通区域分析、灰度一致性分析结合起来用于噪声去除.一方面,我们先用连通区域分析和灰度一致性分析的方法去除二值图片中的一些噪声;另一方面,对于前两种方法不能去除的噪声,基于颜色聚类的方法可以有效去除,从而极大地提高了视频文字的识别结果.在这里,我们使用了现有方法与颜色聚类级联的方式进行噪声去除,这样做的原因是利用这些噪声去除方法的互补性.如果不经过连通分量分析和灰度一致性分析而直接进行聚类去除噪声,那么因为图像中含有较多的噪声,还原到原图上包含的颜色也较多,所以不好确定聚类的类数;同时,也容易把一些噪声和文字笔画聚为一类,不能达到有效去除噪声的目的.在经过了连通分量分析和灰度一致性分析之后,剩下的噪声较少,还原到原图上的颜色比较简单,可以用聚类的方法把噪声和文字笔画分为两类,去掉噪声的一类.下面详细描述基于颜色聚类的去噪方法.

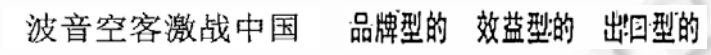


(a) Text block
(a) 文字区域

(b) Binarized result of (a)
(b) 图(a)的二值文字图像

(c) Result of noise removal on (b) using connected component analysis and gray constant constraint
(c) 用连通区域分析和灰度一致性分析对图(b)去除噪声的结果

(d) Colored connected component of (c)
(d) 图(c)中的连通分量添加颜色的结果



(e) Result of noise removal on (d) using color-based clustering
(e) 用颜色聚类的方法对图(d)去除噪声的结果

Fig.9

图 9

对于经过二值化处理结果,我们首先恢复各个连通分量在彩色图片中相应位置的颜色,得到彩色的连通分量,然后用图 10 中的算法进行处理,去除噪声.图 10 中,我们用 *K*-means 的方法对集合 *C* 中的彩色连通分量进行聚类,其中,*k* 的值取为 2,这些连通分量被分为两类.这两类中,所占像素较多的一类被认为是文字,较少一类被认为是噪声.这是因为在经过连通分量分析和灰度一致性分析之后,文字笔画相对与噪声占绝大多数.经过噪声去除的处理,二值图像中的连通分量只剩下文字笔画,或者还可能包含极少量的噪声,这样的图像被输入 OCR 软件进行识别,可以得到较好的识别结果.

1. For each P_i, $P_i \in S$, S is the set of all the connected components
   $c(P_i)=avg(o(R\ G\ B))$, $o \in P_i$
2. $c_{Add}=avg(o'(R\ G\ B))$, $o' \notin P_i$
3. $C=\{c_{Add}\} \cup \{c(P_i)\}$
4. Cluster C into two classes by color: $Class_{text}$ and $Class_{noise}$
5. For each P_i
   If $c(P_i) \in Class_{text}$
   Then P_i is thought to be a text stroke
   Else
   Then P_i is thought to be the noise

Fig.10 Color-Based clustering for noise removal

图 10 本文基于颜色聚类的噪声去除算法

5 实验结果

为了对本文方法进行评价,我们建立了视频帧数据库和图像数据库.其中,视频帧数据库可以对本文的视频文字检测,基于多帧融合视频文字增强和视频文字提取方法进行评测;图像数据库主要对本文的视频文字检测和视频文字提取方法进行评测.视频帧数据库中包含 112 120 个视频帧图像,这些视频帧来自 10 个视频,它们是从多个著名的网站上下载得到,如 CCTV,Xinhuanet 以及 China News 等,分辨率为 320×240.经统计,这些视频帧包含 1 809 行不同内容的文字,每行文字均出现在多个连续的相邻视频帧中,共包含 11 312 个汉字.这些视频帧中的背景往往比较复杂,具有较低的对比度.同时,由于这些视频的压缩率较高,图像质量较差,文字也比较模糊,因此对之进行检测和提取更为困难,可以对本文的方法进行有效评测.实验采用的另一数据库是图像数据库,这个数据库主要用于对本文的视频文字检测模块和视频文字提取模块进行评测,数据库中包含了 2 000 张从网络上随机下载的图片,图片大小在 80×60~800×600 之间.这些图片大多具有复杂的背景,并且都包含文字.这些文字的大小、字体、颜色和对比度各不相同.我们对这 2 000 张图片进行了人工标注,经统计,这 2 000 张图片中总共包含 3 248 行不同内容的文字,文字总数为 22 080.

5.1 视频文字检测

视频文字检测模块的任务是在视频帧中检测并定位文字区域.我们采用了上述的两个实验数据库,用 *recall*(查全率)、*precision*(准确率)和 *f*-measure 等作为文字区域检测结果的评测指标[16].*recall*,*precision* 和 *f* 的取值范围都在 0~1 之间.*recall* 的值越大,说明算法能够检测到相关文字区域的能力越强;*precision* 的值说明了检测到正确文字区域的能力,*precision* 越大,准确率越高;*f* 的值综合考虑了 *recall* 和 *precision*,是对方法总体能力的综合评价.为了与现有方法进行对比,我们实现了 Lyu[10]的方法进行比较,实验结果见表 1.其中,本文方法 I 和本文方法 II 的区别在于,方法 I 没有使用 SVM 方法对文字区域检测结果进行过滤,而方法 II 采用了 SVM 过滤的方法.

Table 1 Compared results of text detection

表 1 视频文字检测方法比较

	Video frame dataset			Image dataset		
	Precision	*Recall*	*f*	*Precision*	*Recall*	*f*
Lyu's method	0.587	0.554	0. 555	0.581	0.747	0.625
Our method I	0.572	0.581	0.558	0.526	0.790	0.592
Our method II	0.615	0.566	0.575	0.620	0.780	0.663

表 1 中,对比本文方法 I 和 Lyu 的方法,本文方法在查全率上得到了提高;但由于引入了一些错误的检测结果,导致了查准率的降低.为了在保持查全率的同时提高查准率,本文用 SVM 的方法对文字区域检测结果进行过滤.对比本文的方法 II 和 Lyu 的方法,本文在 3 个指标上都取得了更好的结果.原因在于:一方面,我们提出的彩色边缘分层方法可以很好地把图像中文字区域的两个显著特征,即密集的文字边缘和一致的文字颜色统一在一起,有利于提高文字区域检测的查全率;另一方面,使用 SVM 过滤也保证了查准率的提高.在这里,我们使用

近邻传播聚类算法对彩色边缘图进行分解,该方法不用预先指定聚类的数目,而由算法根据图像中颜色种类的丰富程度自动决定聚类数目.由于图像中的颜色种类不能预先得到,因此该方法具有更好的普适性.为了比较不同聚类算法分解彩色边缘图得到的文字区域检测结果,对于 *K*-means,我们人为地指定 *k* 的值,使其取值在 2~10 之间,*K*-means 把彩色边缘图分解为 *k* 个子图,并在各个子图上分别检测文字区域.在视频帧数据库上,当 *k* 的值取为 6 时,得到 *f*-measure 的最大值 0.526;在图像数据库上,当 *k* 的取值为 6 时,得到 *f*-measure 的最大值 0.556.与表 1 中的结果比较可以看到,使用近邻传播聚类算法能够取得更好的结果.此外我们也观察到,对在不同测试集上得到的结果进行比较,本文的方法在图像上提高的幅度更大.这是因为图像中文字的颜色比视频中更为丰富,本文的方法更为有效.图 11 给出了一个文字区域检测的例子.图 11(a)是原图,图中有两种颜色的文字.图 11(b)是图 11(a)对应的彩色边缘图,图 11(c)和图 11(d)是对原图进行彩色边缘分层的结果.可以看到,不同颜色的边缘被分解到了不同的边缘子图中去,在各个边缘子图上进行文字检测更加容易和准确.图 11(e)是对(a)进行文字区域检测的结果.

(a) Image (a) 原图 (b) Color edge map (b) 原图对应的彩色边缘图 (c) Decomposed results (c) 彩色边缘分层结果 (d) Decomposed results (d) 彩色边缘分层结果 (e) Result of text detection (e) 文字区域检测结果

Fig.11

图 11

5.2 基于多帧融合的视频文字增强

本文的视频文字识别方法主要包含 4 个模块:视频文字检测模块、视频文字增强模块、视频文字提取模块和 OCR 软件识别模块.其中,视频文字增强模块基于多帧融合来提高文字笔画的清晰度,其最终目的是为了提高视频文字的识别率,现有方法一般采用最终的文字识别结果来对多帧融合的结果进行评测.因此在本文中,为了对视频文字增强模块,即多帧融合方法进行评测,本文在其他 3 个模块采用了相同的方法.这样,多帧融合方法的性能就可以通过文字识别的结果来进行评测.多帧融合的效果越好,文字识别结果就越好.在实验中,本文采用了文献[10]中的视频文字检测方法和视频文字提取方法,并结合了文献[11]中的灰度一致性分析方法来去除噪声;在 OCR 软件识别模块,本文采用了方正锐思 OCR 软件来识别文字.其中,方正锐思 OCR 软件是一个较为成熟的软件,广泛应用在打印文档的识别上.因为本文的方法和比较的方法都采用相同的方正锐思 OCR 软件进行识别,所以识别结果能够比较本文方法的性能.为了与现有的多帧融合模块进行比较,本文实现了如下 4 种不同的多帧融合方法:(I) 文献[6]中的多帧融合方法;(II) 本文的多帧融合方法,包括两个部分:本文提出的文字区域组检测和文献[6]中的文字区域组融合,该方法选取对比度较高的文字区域进行平均融合;(III) 本文的多帧融合方法,包括两个部分:提出的文字区域组检测方法和文献[7]中的文字区域组融合方法,该方法采取了最小值融合;(IV) 本文的多帧融合,包括本文提出的文字区域组检测、文字区域组过滤和文字区域组融合.把上述不同的多帧融合方法和其他 3 个模块进行组合,对数据集中的视频帧进行处理,可以得到不同的文字识别结果.本文采用了文字识别查全率(W_{recall})、文字识别查准率($W_{precision}$)和文字识别重复率(W_{repeat})来对这些识别结果进行评测,这 3 个指标的定义分别如下:

$$W_{recall}=N_{correct}/N_{groundtruth},\ W_{precision}=N_{allcorrect}/N_{recognized},\ W_{repeat}=N_{repeat}/N_{recognized}.$$

这里,采用 W_{repeat} 是因为视频中的文字可能会在不同视频帧中被重复识别,这个指标可以用来衡量文字被重复识别的频率.在上述 3 个指标当中,多帧融合方法的性能主要取决于 $W_{precision}$ 和 W_{recall},这是因为正确识别出文字比重复识别文字更为重要.在上式中:$N_{allcorrect}$ 是正确识别的文字数目,包括重复识别的文字;$N_{correct}$ 是去掉重复识别的文字后正确识别的文字数目;N_{repeat} 是重复识别的文字数目,$N_{allcorrect}$ 的值等于 $N_{correct}$ 与 N_{repeat} 之

和;$N_{recognized}$是识别出的文字,包括错误识别的文字;$N_{groundtruth}$是视频中包含的所有文字.W_{recall}和$W_{precision}$的值越高,表示文字识别结果越好,即相应的多帧融合方法性能越好.

表 2 给出了实验的结果.

Table 2 Compared results of multiple frame integration

表 2 视频文字多帧融合方法比较

	Video frame dataset		
	$W_{precision}$ (%)	W_{recall} (%)	W_{repeat} (%)
Method in Ref.[6] I	54.12	47.55	6.88
Our method II	59.50	54.82	7.59
Our method III	59.28	53.34	8.36
Our method IV	60.43	57.43	8.01

从结果中可以看出,多帧融合方法在查全率和查准率上都取得了比文献[6]中的方法更好的结果,这是因为我们的方法能够更为准确地识别文字区域组.进一步比较方法 II、方法 III 和方法 IV,方法 IV 比方法 II 和方法 III 的查全率和查准率都要高.方法 IV 的改进主要来自两个方面:(1) 方法 IV 基于本文中提出的文字笔画强度图测量文字笔画的清晰度,只选取具有较为清晰的文字区域图像进行融合,从而避免了模糊文字对于融合结果的影响;(2) 方法 IV 结合了平均融合和最小值融合的优点,能够对背景进行简化,提高对比度;同时获取清晰文字,从而能够得到更好的文字提取和识别的结果.图 12 给出了几个用方法 II、方法 III 和方法 IV 融合得到的结果.其中,图 12(a)用文献[6]中方法对左边的文字区域组进行融合得到的结果,选取对比度较大的图像进行平均融合,图 12(b)是用文献[7]中方法得到的结果,对所有图像进行最小值融合得到的结果,图 12(c)是用本文方法进行融合得到的结果.比较图 12(a)~图 12(c)中的结果对比度更高,更利于文字的提取;比较图 12(b)和图 12(c),图 12(c)中的结果文字笔画更为清晰,而图 12(b)中的文字比较模糊,因此,方法 IV 的结果更利于文字的提取和识别.另外,从表 2 中还可以看出,与现有方法相比,因为我们的方法正确识别了更多的文字,所以重复识别的几率更高.如前所述,多帧融合方法的评价主要取决于查全率和查准率,我们的方法在这两个指标上都取得了比现有方法更好的结果,因此综合考虑,我们的方法取得了优于现有方法的结果.此外,我们也与纯平均融合方法进行了比较,即采用本文方法识别文字区域组,并对组内的所有文字区域图像进行平均得到融合图像.该方法得到的查准率、查全率分别是 59.97%和 55.91%,也低于本文方法的 60.43%和 57.43%.

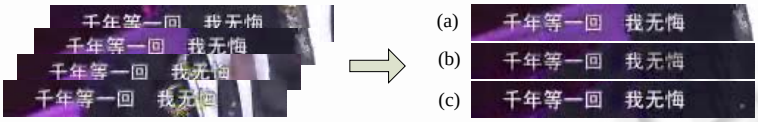


Fig.12 An example of multiple frame integration. Left part is a text block group

图 12 多帧融合结果举例

5.3 视频文字提取

本节对视频文字提取模块进行评测,评测采用了上述的视频帧和图像数据库.为了得到更为准确的评测结果,我们手工标注了这些视频帧和图像中的文字区域.对于在多个视频帧中重复出现的相同文字区域,我们选取其中的 3 幅图像进行评测,包括最先开始的一幅、中间的一幅和最后的一幅.这样,视频帧数据库中包含 1809×3 个文字区域,含有 11312×3 个汉字.在这里,我们之所以对相同的文字区域取不同时间点的 3 幅图像进行评测,是为了证明本文方法处理不同背景的文字区域图像的能力;图像数据库中含有 3 248 个不同的文字区域,总字数为 22 080.在实验中,对于这些手工标注的文字区域,我们用不同的文字提取方法进行处理,把文字区域图像转化为可供 OCR 软件识别的二值文字图像,并输入同样的锐思 OCR 软件进行识别.这样,文字提取模块的性能就可以用文字识别的结果进行评测,评测指标采用了第 5.2 节中定义的 W_{recall}和 $W_{precision}$.为了与现有的方法进行比较,我们实现了如下几种方法:(I) Lyu 的方法[10],该方法利用一个自适应的局部阈值来进行二值化,并且利用连通分量分析的方法去除噪声;(II) 本文的方法,但不包括颜色聚类去噪;(III) 本文的视频文字提取方法.表 3 中,

我们的两种方法在文字识别查全率和查准率上都取得了比 Lyu(方法 I)更好的结果.一方面,这是因为自适应地选取具有较高对比度的颜色分量进行二值化,能够取得较好的二值化结果;另一方面,基于颜色聚类的方法能够进一步去除二值图像中的噪声,改进文字识别的结果.比较方法 II 和方法 III,方法 III 采用了基于颜色聚类的去噪方法,而方法 II 没有.方法 III 所取得的结果好于方法 II,这说明颜色聚类能够有效去除噪声,进一步提高文字识别率.注意到,在方法 III 中,本文采用了级联的噪声去除方式,即首先利用现有方法去除一些噪声,然后再用颜色聚类进一步去噪.这种方式利用了不同噪声去除方法的互补性.如果不经过现有方法的处理而直接进行颜色聚类去噪,因为图像中包含较多的噪声,不易确定聚类的类数,也容易把一些文字笔画和噪声聚为同一类,反而不能得到好的噪声去除效果,甚至可能降低文字识别率.除了上述实验,我们还测试了直接进行颜色聚类去噪的效果,在视频帧数据库上得到的 $W_{precision}$ 和 W_{recall} 值分别是 79.68%和 37.35%;在图像数据库上得到的 $W_{precision}$ 和 W_{recall} 值分别是 81.91%和 49.21%,均不如进行级联去噪得到的结果.

Table 3 Compared results of text extraction

表 3 视频文字提取方法比较

	Video frame dataset		Image dataset	
	$W_{precision}$ (%)	W_{recall} (%)	$W_{precision}$ (%)	W_{recall} (%)
Lyu's method I	78.48	48.27	72.52	53.07
Our method II	79.08	49.86	74.25	58.24
Our method III	83.89	52.71	85.04	68.49

5.4 视频文字识别

前面的实验分别对本文的视频文字检测模块、视频文字增强模块和视频文字提取模块进行了评测.下面对本文方法的整体性能进行测试,测试采用了上述的视频帧数据库.为了与现有的方法进行比较,我们实现了两种方法:

(I) Lyu[10]中的方法,该方法主要包括两个模块:视频文字检测模块和视频文字提取模块;

(II) 本文的方法,包括本文提出的视频文字检测模块、视频文字增强模块和视频文字提取模块.

数据库中的视频帧经过上述两种方法处理之后,得到二值文字图像,把二值文字图像送入同样的锐思 OCR 软件进行识别,并用 W_{recall},$W_{precision}$,和 W_{repeat} 对识别结果进行评测.评测结果如下:Lyu 在 W_{recall},$W_{precision}$,和 W_{repeat} 上的结果分别是 57.07%,40.37%和 4.43%;我们的方法在 W_{recall},$W_{precision}$ 和 W_{repeat} 上的结果分别是 74.05%,63.59%和 12.46%.本文方法在查全率和查准率上都比 Lyu 中方法有较大提高,这是因为本文提出的视频文字检测模块和视频文字提取模块比文献[10]中的相应模块更为有效.同时,基于多帧融合的视频文字增强模块能够有效地利用视频中的冗余文字信息进一步改进文字识别的结果.这些都在前面的实验中得到了证明.

6 总结及展望

本文提出了一种基于颜色聚类和多帧融合的视频文字识别方法,主要具有如下创新之处:

(1) 在视频文字检测模块,充分考虑了文字区域的两个显著特征:密集文字边缘和一致文字颜色,采用近邻传播聚类算法把不同颜色的边缘分解到若干边缘子图中,使得在各个边缘子图中检测和定位文字区域更加准确;

(2) 在视频文字增强模块,对包含相同内容的多个文字区域进行融合,利用文字笔画强度图过滤掉那些笔画较为模糊的文字区域,并综合了平均融合和最小值融合的优点,能够有效地平滑文字背景和增强文字笔画的清晰度;

(3) 在视频文字提取模块,一方面,通过自适应地选取文字对比度较高的颜色分量进行二值化,能够得到更好的二值化结果;另一方面,考虑了图像中文字与背景的颜色差异,采用颜色聚类的方法,并结合连通分量分析和灰度一致性分析,能够更为有效地去除噪声.

在接下来的工作中,我们将进一步研究视频中复杂字幕的识别技术,例如视频中运动字幕的追踪和融合技术,以及排列不规则字幕的检测和提取等;我们还将研究如何利用识别出来的字幕对视频内容进行更为准确的标注.

References:

[1] Liu Y, Goto S, Ikenaga T. A robust algorithm for text detection in color images. In: Proc. of the 8th Int'l Conf. on Document Analysis and Recognition. 2005. 399−403. [doi: 10.1109/ICDAR.2005.29]

[2] Karatzas D, Antonacopoulos A. Text extraction from Web images based on a split-and-merge segmentation method using colour perception. In: Proc. of the 17th Int'l Conf. on Pattern Recognition. 2004. 634−637. [doi: 10.1109/ICPR.2004.1334328]

[3] Gllavata J, Ewerth R, Freisleben B. Text detection in images based on unsupervised classification of high-frequency wavelet coefficients. In: Proc. of the 17th Int'l Conf. on Pattern Recognition. 2004. 425−428. [doi: 10.1109/ICPR.2004.1334146]

[4] Zhuang YT, Liu JW, Wu F, Pan YH, Zhang Y. Automatic caption location and extraction in digital video based on support vector machine. Journal of Computer Aided Design and Computer Graphics, 2002,14:750−753 (in Chinese with English abstract).

[5] Li HP, Doermann D. Text enhancement in digital video using multiple frame integration. In: Proc. of the 7th ACM Int'l Conf. on Multimedia. 1999. 19−22. [doi: 10.1145/319463.319466]

[6] Hua XS, Yin P, Zhang HJ. Efficient video text recognition using multiple frame integration. In: Proc. of the 9th Int'l Conf. on Image Processing. 2002. 397−400. [doi: 10.1109/ICIP.2002.1039971]

[7] Lienhart R, Wernicke A. Localizing and segmenting text in images and videos. IEEE Trans. on CSVT, 2002,12(4):256−268. [doi: 10.1109/76.999203]

[8] Sato T, Kanade T, Hughes E, Smith M, Satoh S. Video OCR: Indexing digital news libraries by recognition of superimposed captions. Multimedia Systems, 1999,7(5):385−385. [doi: 10.1007/s005300050140]

[9] Mi CJ, Liu Y, Xue XY. Video texts tracking and segmentation based on multiple frames. Journal of Computer Research and Development, 2006,43:1523−1529 (in Chinese with English abstract).

[10] Lyu MR, Song JQ, Cai M. A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Trans. on CSVT, 2005,15(2):243−255. [doi: 10.1109/TCSVT.2004.841653]

[11] Chen DT, Odobez JM, Bourlard H. Text detection and recognition in images and video frames. In: Proc. of the Pattern Recognition. 2004. 595−608. [doi: 10.1016/j.patcog.2003.06.001]

[12] Zhan YW, Wang WQ, Gao W. A robust split-and-merge text segmentation approach for images. In: Proc. of the 18th Int'l Conf. on Pattern Recognition. 2006. 1002−1005. [doi: 10.1109/ICPR.2006.169]

[13] Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007,315(5814):972−976. [doi: 10.1126/science.1136800]

[14] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans. on Systems, Man and Cybernet, 1979:62−66. [doi: 10.1109/TSMC.1979.4310076]

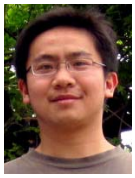
[15] Wolf C, Jolion JM. Extraction and recognition of artificial text in multimedia documents. In: Proc. of the Pattern Analysis and Application. 2003. 309−326. [doi: 10.1007/978-3-642-13318-3_37]

[16] Lucas SM, Panaretos A, *et al*. ICDAR 2003 robust reading competitions. In: Proc. of the 7th Int'l Conf. on Document Analysis and Recognition. 2003. 682−687. [doi: 10.1109/ICDAR.2003.1227749]

附中文参考文献:

[4] 庄越挺,刘骏伟,吴飞,潘云鹤,张引.基于支持向量机的视频字幕自动定位与提取.计算机辅助设计与图形学学报,2002,14(8):750−753.

[9] 密聪杰,刘洋,薛向阳.基于多帧图像的视频文字跟踪和分割算法.计算机研究与发展,2006,43(9):1523−1529.



易剑(1984－),男,贵州晴隆人,博士生,主要研究领域为多媒体信息处理.

彭宇新(1974－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为图像、视频内容理解与检索,网络多媒体内容的搜索与挖掘,网络多媒体内容安全.

肖建国(1957－),男,教授,博士生导师,CCF 高级会员,主要研究领域为图像,影像处理,网络信息处理,文本挖掘技术.