一种基于多帧视频的文本图像质量增强方法

朱成军 李 超 薛 玲 熊 璋
(北京航空航天大学计算机学院 计算机应用研究室,北京 100083)

摘 要 视频文本和视频内容高度相关,提供了理解视频内容的有用信息,然而文本往往位于复杂背景之中,从视频帧中定位到文本区域后,如果将其直接送入 OCR 软件,其识别效果较差。视频文本的时域信息提供了增强文本,消除背景的有用信息。因此,提出了一种利用视频文本的时域信息来消除背景,增强文本的方法。该方法首先利用边缘算子计算文本的轮廓特征,然后采用基于 Hausdorff 距离度量的匹配方法跟踪本文区域在相邻帧序列中的位置,利用多帧平均或帧间最小搜索法消去背景;其次,利用双线性插值技术调整文本尺寸,最终得到具有干净背景、合理分辨率的文本图像。不同测试视频序列的实验结果表明,该方法可以有效提高视频文本的 OCR 软件识别率。

关键词 视频分析 文本追踪 文本增强 Hausdorff 距离
中图法分类号:TP391 **文献标识码**:A **文章编号**:1006-8961(2008)09-1667-06

Video Text Enhancement Using Multiple Frame Information

ZHU Cheng-jun, LI Chao, XUE Ling, XIONG Zhang
(*School of Computer Science and Technology*, *Beihang University*, *Beijing* 100083)

Abstract Text in video is a very compact and accurate clue for video indexing and summarization. But video texts are usually embedded in complex background, making it very difficultl for text separation from the background information. Hence the OCR accuracy was poor. This paper presents a multi-frames based technique to enhance video text image. After extracting a reference text block, we use Hausdorff distance based image matching technique to find and register the corresponding text block. Then the frames average or minimum pixel search method is applied to text blocks to obtain a new text block with a clean background. At last we apply a finite interpolation function to adjust the text block resolution. Experiments conducted on several video sequences show that our enhancement scheme can considerably improve the accuracy of OCR.

Keywords video analysis, text tracking, text enhancement, Hausdorff distance

1 引 言

视频中的文本提供了和视频内容高度相关的信息,比如场景地点、事件时间,以及体育比赛中的比分、运动员姓名等信息,但是相对于文档图像中的文本,视频中的文本识别面临以下难点:(1)由于电视制式、视频传输和存储的原因,视频图像分辨率较低;(2)视频中的文本往往叠加在视频场景中。由于大多数商用 OCR 软件只能处理具有干净背景的二值图像,所以在检测到视频中的文本区域以后,还必须将文本和背景分离开来,得到具有干净背景的文本图像,然后才能进行 OCR 识别。

关于图像二值化,已经有了很多的相关工作[1~3],但是这些方法并不适用于视频文本的二值化操作。图 1(a)是一幅视频帧中的文本区域图像,

基金项目:航空支撑科技基金项目(05E551010)
收稿日期:2007-01-25;**改回日期**:2007-04-09
第一作者简介:朱成军(1978~),男。现为北京航空航天大学计算机应用技术专业博士研究生。主要研究方向为多媒体检索技术、图像处理、模式识别。E-mail:zhuchengjun@163.com

图 1(b)和图 1(c)是采用 Otsu[1]得到的二值图像及相应的反色结果,可以看出,背景和文本混杂在一起,如果将图 1(c)直接输入到 OCR 软件,文字识别率并不是很理想。现有的视频文本质量增强研究工作,主要有基于单帧[4,5]和多帧[6]两种方法。基于单帧的方法利用文本的连通性和水平分布特征,采用形态学和连通域的分析方法,在简单背景下,可以很好地将背景和文本分离,但是当文本具有稍微复杂的背景时,该方法效果不是特别理想。

(a)视频中的文本区域

(b)二值化的结果

(c)颜色反转的结果

图 1 视频文本区域二值化示例

Fig. 1 Binarization of text region detected in video

利用多帧信息在时域方向的特点消除背景,关键在于追踪文本在相邻帧的准确位置。关于视频对象追踪,已有很多的研究成果[7],比如人脸追踪、汽车追踪等,但是很少的研究涉及视频中文本的追踪,文献[6]提出了一种利用帧间差平方和(SSD)追踪视频文本的方法,但是该方法在文本具有运动背景的时候准确率不是太高。

综合起来,视频文本区域在时域具有以下特点:(1)同一文本会持续几十帧甚至几百帧;(2)当文本有运动时,呈现水平或者垂直的线型运动;(3)文字的生存期,文字像素颜色保持不变,背景像素颜色可能具有很大的变化。所以,如果能将多帧的文本区域信息综合起来,生成新的文本图像,动态变换的背景像素会得到削弱。

基于视频文本区域的上述时域特征,本文的文本增强算法如图 2 所示,首先利用文本检测算法检测新出现的文本区域,然后利用边缘算子提取其边缘特征,以其作为初始模板,使用 Hausdorff 距离度量方法跟踪文本在相邻帧的位置,利用多帧平均或最小像素搜索法消除背景,最后通过线性插值调整文本图像分辨率,得到具有干净背景、合理分辨率的文本图像。

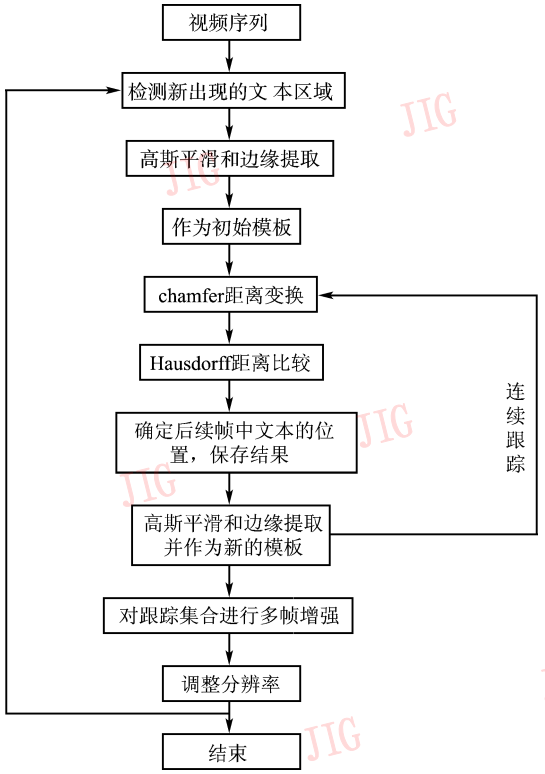


图 2 视频文本跟踪及增强的原理图

Fig. 2 Video text tracking and enhancement paradigm

2 文本检测

基于多帧的文本增强算法,第 1 步需要检测文本出现的起始位置。由于视频的帧率在 30fps 左右,如图 2 所示,对视频帧集合 $\boldsymbol{C}$,视频帧 $f_t \in \boldsymbol{C}$,利用文本检测算法每隔 15 帧检测新出现的文本区域 $r_i(f_t)$。关于文本检测,很多研究人员做了大量的研究工作[8~10]。文献[11]基于文本的边缘特征和空间分布特征,利用支持向量机(SVM)检测文本区域 $r_i(f_t)$。

3 文本追踪

3.1 文本特征提取

文本区别于背景的特征包括连通性和对比度 2 种。连通性特征是指视频中的文本笔画按水平或者垂直方向有规律分布,在出现周期具有一致的颜色,而对比度特征是指为了便于观众阅读,文本像素和背景像素颜色值具有较大的差异值。文献[6]利用文本的连通性特征,计算当前文本区域与下一帧相

邻区域的 SSD 值,使用帧间 SSD 值作为匹配的准则。跟踪过程描述如下:给定视频帧 $\boldsymbol{I}$ 中的参考文本区域,然后在视频帧 $\boldsymbol{J}$ 中的给定区域 $\boldsymbol{W}$ 中进行搜索,返回具有最小 SSD 值的位置。定义如下:

$$\varepsilon = \iint_{W} [\boldsymbol{J}(\boldsymbol{D}\bar{\boldsymbol{x}} + \bar{\boldsymbol{d}}) - \boldsymbol{I}(\bar{\boldsymbol{x}})]^2 \mathrm{d}\bar{\boldsymbol{x}} \tag{1}$$

式中,$\boldsymbol{D}$ 是 2×2 的形变矩阵,$\bar{\boldsymbol{d}}$ 是位移矢量。但是当文本具有运动背景时或者具有噪声时,*SSD* 会具有较大的值,所以将 *SSD* 作为匹配准则,并不能很好地适应背景和文本运动的情形。

根据人视觉系统的认知特点,文字区别于非文字对象,或者不同于文字之间的主要区别,在于文字笔画的不同结构组合,也就是说,某个文字对象的结构是独一无二的。如图 3 所示,在相邻帧中,视频中的文本亮度由暗到比较明亮,帧间文本块具有较大的 *SSD* 值,然而观察文本在不同帧上的边缘特征,相似度是最高的。所以选择文本的边缘作为帧间匹配的判断准则。

(a)视频流第45帧的文本区域和对应边缘图像

(b)视频流第47帧的文本区域和对应边缘图像

(c)视频流第49帧的文本区域和对应边缘图像

图 3　文本区域在不同帧的边缘特征

Fig. 3　Edge feature of video text in different video frames

文本的边缘特征提取步骤如下:

(1)使用高斯滤波器对文本块进行平滑,减少噪声影响。

(2)计算每一点的梯度方向 $\theta(x,y)$ 以及局部梯度值 $F_\theta(x,y)$。

$$\theta(x,y) = \frac{1}{2}\arctan\left[\frac{2g_{xy}}{(g_{xx}+g_{yy})}\right] \tag{2}$$

$$F_\theta(x,y) = \left\{\frac{1}{2}[(g_{xx}+g_{yy}) + (g_{xx}+g_{yy})\cos 2\theta + 2g_{xy}\sin 2\theta]\right\}^{\frac{1}{2}}$$

其中,g_{xy}, g_{xx}、g_{yy}为 x,y 方向的梯度点积。

(3)基于 $\theta(x,y)$ 和 $F_\theta(x,y)$,采用 Canny 算子得到文本的连续边缘。

3.2　Hausdorff 距离度量

给定两组有限点集合 $\boldsymbol{A} = \{a_1, \cdots, a_p\}$ 和 $\boldsymbol{B} = \{b_1, \cdots, b_q\}$,Hausdorff 距离定义为

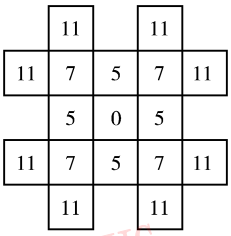
$$H(\boldsymbol{A},\boldsymbol{B}) = \max(h(\boldsymbol{A},\boldsymbol{B}), h(\boldsymbol{B},\boldsymbol{A})) \tag{3}$$

其中,

$$h(\boldsymbol{A},\boldsymbol{B}) = \max_{a\in A}\min_{b\in B}\| a - b \| \tag{4}$$

$$h(\boldsymbol{B},\boldsymbol{A}) = \max_{b\in B}\min_{a\in A}\| b - a \|$$

函数 $h(\boldsymbol{A},\boldsymbol{B})$ 称为从 $\boldsymbol{A}$ 到 $\boldsymbol{B}$ 的有向 Hausdorff 距离,其意义是在集合 $\boldsymbol{A}$ 的任意一点 $a\in\boldsymbol{A}$ 到集合 $\boldsymbol{B}$ 中所有点距离中选出最近距离,再考虑在集合 $\boldsymbol{A}$ 的每一点到集合 $\boldsymbol{B}$ 最近距离集合中选择最大值。$h(\boldsymbol{B},\boldsymbol{A})$的意义同 $h(\boldsymbol{A},\boldsymbol{B})$相似,称为反向 Hausdorff 距离。在实际计算 Hausdorff 距离时,是将已知模板二值图像与将匹配的未知二值图像转换为距离函数。距离函数是将二值图像转换为另一种灰度图像,二值图像的“1”,对应于距离图像的“0”。而二值图像的“0”,依据图 4(a)的距离变换系数转换为不同的值,在距离图像中离“0”越近,距离值越小,反之会很大。如图 4(b)、图 4(c)所示。



(a)距离变换系数chamfer5-7-11

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	1	1	1	1	0	0
0	0	1	0	0	1	0	0
0	0	1	0	0	1	0	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

(b)二值图像

14	11	10	10	10	10	11	14
11	7	5	5	5	5	7	11
10	5	0	0	0	0	5	10
10	5	0	5	5	0	5	10
10	5	0	5	5	0	5	10
10	5	0	0	0	0	5	10
11	7	5	5	5	5	7	11
14	11	10	10	10	10	11	14

(c)距离变换结果

图 4　Hausdorff 距离变换

Fig. 4　Transformation based on Hausdorff distance

3.3 基于 Hausdorff 距离度量的跟踪算法实现

当在下一帧中匹配当前文本时,如果搜索整个视频帧,算法时间复杂度会很高,也没有必要,可通过估计文本的最大运动速度来缩小匹配范围。由于文本需要辅助观众理解视频内容,所以文本不可能运动速度太快。由于网络带宽和存储的原因,现在的视频分辨率大多限制在 352×240 以下,帧率大约在 20~30fps 之间。如果文本的出现时间为 δt,那么当前文本在相邻帧垂直和水平方向上的偏移值(以像素为单位)为

$$\delta v = \frac{240}{20 \times \delta t} \quad \delta h = \frac{352}{20 \times \delta t} \tag{5}$$

根据视频中的文本时间滞留时间统计,设置 δt 的最小估计值为 3s,所以根据式(3),在相邻帧,文本的最大偏移值 $\delta v \approx 4$ 像素、$\delta h \approx 6$ 像素。

如果当前文本区 $r_i(f_t)$ 的矩形面积为 $w_i \times h_i$,根据最大偏移值 δv 和 δh,可以估计出在相邻下一帧的匹配区域面积大小为 $(w_i + 2 \times \delta h) \times (h_i + 2 \times \delta v)$,具体的跟踪算法如下:

(1)对于文本区域 $r_i(f_t)$,用 3.1 节介绍的基于彩色的边缘轮廓提取方法,生成边缘二值图像,作为 Hausdorff 距离比较的集合 $\boldsymbol{B}$;

(2)根据式(5)的方法计算相邻下一帧的搜索区域,并采取和步骤 1 同样的方法生成其缘二值图像,作为 Hausdorff 距离比较的集合 $\boldsymbol{A}$;

(3)用图 4(a)所示的距离变换系数,将二值图像 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 变换为距离图像 $\boldsymbol{D}'$和 $\boldsymbol{D}$;

(4)逐点扫描图像 $\boldsymbol{A}$,并对其任一点(x,y)进行计算,在对图像 $\boldsymbol{B}$ 中所有“1”的像素点(k,l),查找相应 $\boldsymbol{A}$ 的距离图像$(x'+k,y'+l)$点的值,放入数组 ***Sort*** 中;

(5)计算数组 ***Sort*** 中所有值的平均值,作为点(x',y')的有向 Hausdorff 距离值,放入数组 $\boldsymbol{F}_B$;

(6)考虑在点(x',y')处,图像 $\boldsymbol{B}$ 覆盖图像 $\boldsymbol{A}$ 的区域,对图像 $\boldsymbol{A}$ 在这个区域内所有“1”的像素点(k_1, l_1),查找相应距离图像 $\boldsymbol{B}$ 的$(k_1 + x', l_1 + y')$点的值,放入数组 $\boldsymbol{Sort}_1$ 中;

(7)计算数组 $\boldsymbol{Sort}_1$ 中所有值的平均值,并将其作为点(x',y')反向 Hausdorff 距离值,放入数组 $\boldsymbol{F}_A$ 中;

(8)比较在点(x',y')上,$\boldsymbol{F}_A$ 和 $\boldsymbol{F}_B$ 的大小,将较大的值放入数组 $\boldsymbol{F}$ 中;

(9)返回第 4 步,计算图像 $\boldsymbol{A}$ 所有点的 $\boldsymbol{F}$ 值;

(10)查找数组 $\boldsymbol{F}$ 中的最小值 min($\boldsymbol{F}$),根据式(6)计算差异值。

$$\varepsilon = \frac{\min(F)}{w_i \times h_i} \tag{6}$$

当 ε 小于给定阈值,则该点作为匹配对象的位置,如果 ε 大于阈值,该文本的追踪结束;

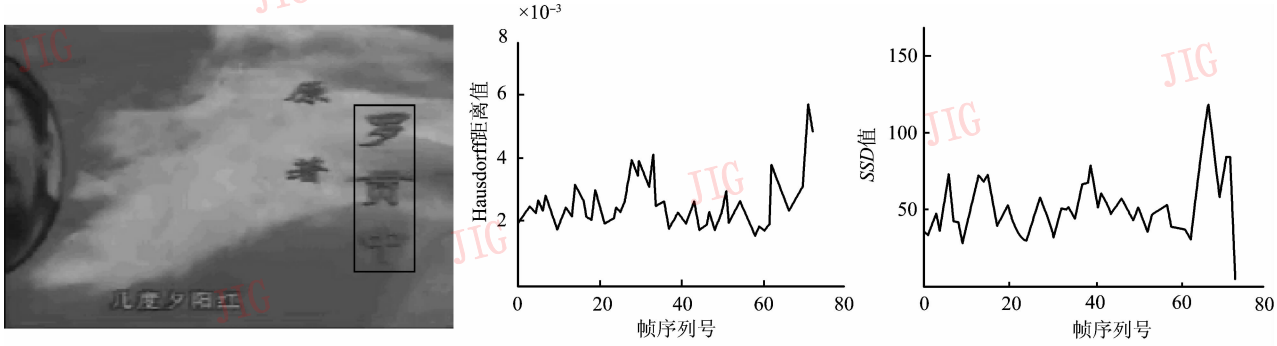
(11)保存 $\boldsymbol{A}$ 中得到的匹配结果,并用其更新 $r_i(f_t)$,作为新的模板来匹配下一帧中的文本,返回步骤 1。

图 5 给出了一个跟踪结果,其中,图 5(a)黑色方框内为视频序列中文本出现的初始区域,该视频序列的背景和文本都是运动的,文本从右向左线性运动。采用本文的跟踪算法,在文本消失时,如图 5(b)所示,得到一个很大的距离值,然而文献[6]的 *SSD* 峰值较为提前。

4 文本质量增强

4.1 背景消除

本算法将视频帧中像素划分为两种类型,文本



(a)视频中的文本区域 (b)帧间Hausdorff距离值 (c)帧间*SSD*值

图 5 视频序列中的文本跟踪结果

Fig. 5 Tracking result of text in video frame sequences

像素和背景像素。当文本出现在相邻帧中时,文本像素的颜色是保持不变的,然而当背景是运动的时候,背景像素的颜色值会发生变化。主要有两类基于多帧的文本增强技术(最小值像素搜索[8]和帧平均法[9])可用于减少背景干扰,增强文本质量。根据文本追踪结果,令 $\boldsymbol{C}_i$ 视频帧集合,帧 $f_t \in \boldsymbol{C}_i$ 包含同一个文本区域 $r_i(f_t)$。

4.1.1　多帧平均法

如果将所有同一文本区域的图像进行像素值求和并进行平均,可以得到一个新的文本图像,在新文本图像中,文本的像素值保持不变,而背景得到了平滑,从而更加容易分离文本和背景,得到文本的二值图像用于识别。

$$\bar{\gamma} = \frac{1}{|\boldsymbol{C}_i|}\sum_{f_t \in \boldsymbol{C}_i}\gamma_i(\boldsymbol{f}_i) \tag{7}$$

式中,$|\boldsymbol{C}_i|$表示视频帧的数量。

4.1.2　最小像素搜索法

将每个像素灰度值在时间维的值看作 1 维向量,由于文本的灰度值是稳定的,而背景是变化的,在时间方向上取每个像素值的最小值,可以减少背景的影响。

$$\hat{\gamma} = \min_{f_t \in \boldsymbol{C}_i}\gamma_i(\boldsymbol{f}_i) \tag{8}$$

4.2　调整分辨率

对于文档图像,一般每英寸包含 300 像素,字符区域为 20×20 ~ 40×40 像素大小面积,此分辨率可以保证 OCR 软件具有良好的识别率。但是视频中的文本具有不同大小的字体,字符大小从最低 10×10往上增加不等。为了提高视频文本的识别率,在送入 OCR 软件前,需要对文本图像进行增加或者降低图像分辨率的插值操作。

插值操作可以看成重采样的过程,假定有离散的低分辨率文本图像 $\boldsymbol{f}$,利用 sinc 函数可以恢复连续的文本图像 $\boldsymbol{f}_t$,表示如下:

$$\boldsymbol{f}_t = \boldsymbol{f} \otimes \mathrm{sinc}(t) \tag{9}$$

式中,t 是连续时间变量。期望的高分辨率离散图像 $\boldsymbol{f}'$ 可采用更高的采样率 t_s 经由 $\boldsymbol{f}_t$ 重采样得到:

$$\boldsymbol{f}' = \boldsymbol{f}_t\sum_{m=-\infty}^{m=+\infty}\delta(t - mt_s) \tag{10}$$

sinc 函数是理想低通滤波的冲击响应,具有无限长的支持域,所以式(9)、式(10)的计算复杂度较高,这为 sinc 函数插值算法的实现带来了困难。因此,本文采用有限支持域的双线性插值函数,虽然双线性插值函数没有 sinc 函数一样的低通滤波性,但是在插值效果和计算速度之间是一个很好的折中选择。

5　实验结果

对视频中背景和文本的以下 4 种组合情况进行测试:(1)文本静止、背景静止;(2)文本静止、背景运动;(3)文本运动,背景静止;(4)文本运动、背景运动。测试数据包括 6 个视频序列,文本的识别采用 Microsoft office document imaging 中的 OCR 模块。图 6 和图 7 是视频序列中具有剧烈运动和轻微运动背景的两个文本区域的识别结果,未识别出的文字以“?”代替。每一个文本区域分别是采用帧平均法和最小像素搜索法进行质量增强,从识别结果来看,识别的文字已完全能表达整句话意义。

图 6　文本多帧增强结果(运动、复杂背景)

Fig. 6　Result of video text enhancement(complex moving background)

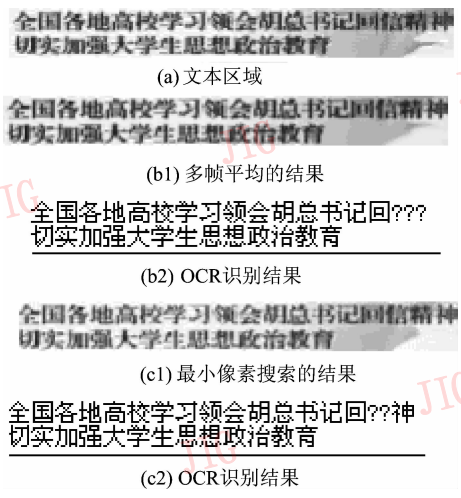


图 7　文本多帧增强结果(静止、简单背景)

Fig. 7　Result of video text enhancement(simple static background)

表 1 给出了进一步的实验结果,可以看出,当文本或者背景具有运动时,文本序列在时域上的信息加强了文本像素,并且提供了多余的信息消除背景像素的影响,从而达到了增强视频文本的效果。对于背景和文本均静止的情形,由于在时域方向并没

有冗余的信息用于区分文本和背景像素,质量增强算法的效果并不明显。根据视频文本存在的目的(辅助观众理解视频的内容),结合测试视频的内容,可以得知,这样的结果是合理的。为了保持和背景的高对比度,便于观众阅读,当文本和背景保持静止时,背景不能太复杂。而当文本位于复杂背景时,背景必然是运动的。这和观众平常观看电影、体育节目和新闻节目时的大体感受是一致的。

表 1 OCR 识别结果

Tab. 1 OCR recognition result

描述(数量)	背景、文本	帧数	字符	单帧 OCR 识别数/识别率(%)	多帧 OCR 识别数/识别率(%)
电视片头(2)	运动、运动	133	80	31/38.8	66/82.5
电视片头	静止、静止	69	62	45/72.6	47/75.8
电影片头(2)	运动、静止	146	86	50/58.1	76/86.4
新闻	静止、运动	50	69	39/56.5	54/78.2
共计		398	288	165/57.3	243/84.4

6 结 论

视频中的文本和当前视频内容高度相关,但是和文档图像中的文本不一样,视频文本往往位于复杂背景中,但由于视频在时域包含了冗余和互补的信息,针对文本在时域中的这个特点,提出并实现了一种基于文本对象跟踪,利用多帧信息的文本质量增强方法。实验结果表明,在文本或者背景运动的情形下,质量增强算法可以显著提高 OCR 软件识别的准确率。

参考文献(References)

1 Otsu N A. Threshold selection method from grey-level histograms [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, **9**(1): 377 ~ 393.

2 Mardia K V. Hainsworth T J. A spatial threshold method for image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, **10**(6): 919 ~ 927.

3 Niblack W. An Introduction to Digital Image Processing [M]. Englewood Cliffs: Prentice Hall International, 1986:115 ~ 116.

4 Tang X, Gao X, Liu J. A spatial-temporal approach for video caption detection and recognition [J]. IEEE Transactions on Neural Networks, 2002, **13**(4):961 ~ 971.

5 Lyu M R, Song Ji-qiang, Cai Min. A comprehensive method for multilingual video text detection, localization, and extraction [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, **15**(2):243 ~ 255.

6 Li Hui-ping, Doermann D. Text enhancement in digital video using multiple frame integration [A]. In: Proceedings of ACM Multimedia [C], Orlando FL, USA, 1999:19 ~ 22.

7 Han Jun, Xiong Zhang, Sun Wen-yan. A method implementation of automatic segmenting and tracking of video moving objects [J]. Journal of Image and Graphics, 2001, **6**(8):732 ~ 737. [韩军, 熊璋, 孙文彦. 自动分割及跟踪视频运动对象的一种实现方法 [J]. 中国图象图形学报, 2001, **6**(8): 732 ~ 737]

8 Li H, Doermann D S. Automatic identification of text in digital video key frames [A]. In: Proceedings of 14th International Conference on Pattern Recognition[C], Piscataway, NJ, USA, 1998:129 ~ 132.

9 Sato T, Kanade T, Kughes E K, *et al*. Video OCR: indexing digital news libraries by recognition of superimposed captions [J]. ACM Multimedia, 1999, **7**(5):385 ~ 395.

10 Li H P, Doemann D, Kia O. Text extraction, enhancement and OCR in digital video [A]. In: Proceeding of 3rd IAPR Workshop [C], Nagoya, Japan, 1998: 363 ~ 377.

11 Zhu Cheng-jun, Ouyang Yuan-xin, Gao Lei. An automatic video text detection, Localization and extraction approach[A]. In: Proceedings of the 2006 Conference on Signal-Image Technology & Internet-based Systems[C], Hammamet, Tunisia, 2006:166 ~ 175.