一种视频文本自动定位、跟踪和识别的方法

李朝晖[1),2)] 余英林[2)]

[1)](广州大学信息学院计算机系,广州 510405) [2)](华南理工大学电子与通信工程系,广州 510641)

摘 要 视频数据中的文本能提供重要的语义信息。本文提出了一种视频文本自动定位、跟踪和识别的方法,首先用基于小波和 LH 检测视频帧文本所在的位置,然后用运动估计的方法,跟踪后继帧文本的位置,再用多帧平均的方法增强文本区域,最后经过二值化处理和连通分量分析,将文本字符送入 OCR 软件进行识别。实验结果表明,该方法简单易行,能快速地定位和跟踪文本区域,定位精度和识别效果良好。

关键词 文本检测 语义内容 视频索引

中图法分类号: TP391.41 **文献标识码**: A **文章编号**: 1006-8961(2005)04-0457-06

An Algorithm of Automatic Video Text Locating, Tracking and Recognition

LI Zhao-hui[1),2)], YU Ying-lin[2)]

[1)](*Department of computer, Information School, Guangzhou University, Guangzhou* 510405)

[2)](*Department of Communication and Electronic Engineering, South China University of Technology, Guangzhou* 510641)

Abstract Text in video can provide an important supplemental source of index semantic information. In this paper, an algorithm of automatic video text locating, tracking and recognition is presented. First, the text regions are located by several steps: wavelet decomposition, high frequency component intensity and density detection, horizontal and vertical convex detection based LH, and text locating. Then the text regions are tracked in next consecutive frames. After multiple frames averaging, the text regions are enhanced. By binarization of the enhanced text regions followed by component analysis, the text regions with clean background are obtained. Then the text regions are recognized by OCR software, the final text strings are attained. Experimental results show that the proposed algorithm can detect and track text region simply and effectively.

Keywords text detecting, semantic context, video index

1 引 言

视频中的文本信息如视频中的片名、人物对话、演员表等能提供重要的语义信息。而现有的 OCR 识别系统还不能直接识别复杂背景下的文本,因而,从视频中提取和跟踪文本区域,具有重要的实际意义。

近年来,已有不少学者提出了提取视频文本的相关算法[1~5]。如使用一个混合的小波/神经网络对 16×16 的像素块进行检测[1];利用文本图像在水平和垂直方向亮度的有规律变化在 8×8 的 DCT 压缩域直接实现文本检测[2,3],这种基于块的方法,其主要缺点是定位边界不精确。另外,从现有的文献资料来看,大多数学者都限于对视频文本的检测,而实现文本跟踪和字符识别的相应比较少。现有的方法虽能在一定程度上解决文本检测的问题,但并不完善。主要难点在于:视频帧本身的低分辨率、视频文本所处背景的复杂性,文本的颜色、字体大小的可变性等使得文本区域的检测仍然是难度较大的问题。因此,如何精确地检测和跟踪文本边界、简化背景,仍然是个开放的问题,有着非常重要的意义。

本文提出一种视频文本自动定位、跟踪和识别

基金项目:国家自然科学基金项目(60372068);广东省科学基金项目(011628)

收稿日期:2004-01-08;**改回日期**:2004-09-06

第一作者简介:李朝晖(1966~),女,副教授。2004 年毕业于华南理工大学,获通信与信息专业博士学位。主要研究方向为图像和视频分析、模式识别。E-mail:lichaoh@163.net

的方法，利用文本信息丰富的纹理特征，尝试用具有良好时频局部特性和变尺度特性的小波分析方法，提取纹理清晰、具有不同空间分辨率、不同方向的边缘子图像，经过强度、密度检测后，用 LH 的方法对提取出的细节图像进一步定位文本目标区域，然后用运动估计的方法，跟踪后继帧文本的位置，再用多帧平均的方法增强文本区域，最后经过二值化处理和连通分量分析，将文本字符送入 OCR 软件进行识别。实验证明，该方法简单易行，能快速地定位和跟踪文本区域，定位精度和识别效果良好。

2 算法原理

观察视频文本发现，视频文本一般具有比较丰富的纹理信息。因此，其高频分量比文本本身的颜色具有更可靠的特征。此外，文本具有一定的空间黏附性，即同一文本行具有相似的高度、方向和距离。正是基于以上特征，设计了一个简单有效的方法检测复杂背景下的文本区域。该方法的处理模块如图 1 所示。

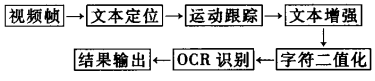

图 1 方法的处理模块

Fig. 1 The processing modules of the algorithm

2.1 文本定位

文本定位的原理框图如图 2 所示，利用文本信息丰富的纹理特征，将视频帧用小波分解后得到的边缘子图像，经过强度、密度检测后，用 LH 的方法对提取出的细节图像进一步定位文本目标区域。

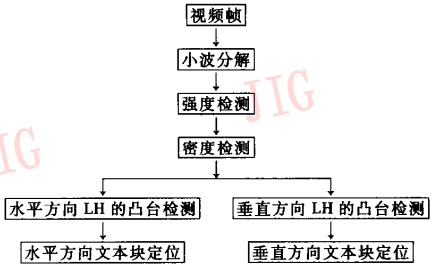

图 2 文本定位模块

Fig. 2 The text locating module

2.1.1 小波分解

小波变换能够把时间域、频率域和空间域有机地结合起来，其最大特点是能够对信号进行显微镜式的观察，对高频信号进行细处理，而对低频信号进行粗处理。由于正交小波基的 Haar 小波变换，能在水平、垂直、对角方向进行分解，因而它更适合于文字区域的检测（比如中文的最基本笔画是横、竖、捺、撇，分别对应水平、垂直、对角方向），所以选取 Haar 小波对视频帧进行分解。

2 维的 Haar 小波分解算法如下：

$$LL_{x,y}=\frac{1}{4}(I_{2x,2y}+I_{2x,2y+1}+I_{2x+1,2y}+I_{2x+1,2y+1})$$

$$LH_{x,y}=\frac{1}{4}(I_{2x,2y}-I_{2x,2y+1}+I_{2x+1,2y}-I_{2x+1,2y+1})$$

$$HL_{x,y}=\frac{1}{4}(I_{2x,2y}+I_{2x,2y+1}-I_{2x+1,2y}-I_{2x+1,2y+1})$$

$$HH_{x,y}=\frac{1}{4}(I_{2x,2y}-I_{2x,2y+1}-I_{2x+1,2y}+I_{2x+1,2y+1})$$

其中，$LL_{x,y}$，$LH_{x,y}$，$HL_{x,y}$，$HH_{x,y}$ 分别代表近似、垂直、水平和对角分量，x，y 代表坐标，I 代表亮度。

从视频序列中选取一帧彩色图（图版 Ⅰ 图 1），按如下的公式将其转化为灰度图：

$$Y=0.299R+0.587G+0.114B$$

这里，Y 代表灰度，R、G、B 分别代表彩色图的红绿蓝分量。

将灰度图进行小波分解，图版 Ⅰ 图 2 是其一级小波分解图，上排从左到右分别为近似、水平分量，下排从左到右分别是垂直以及对角方向的分量图。从图版 Ⅰ 图 2 可以看出，小波分解后的图中有能较好地体现文本位置的信息，特别是水平、垂直以及对角方向高频分量图。

将水平、垂直以及对角方向分量图按一定的权值组合成一新的分量图

$$E=\alpha H+\beta V+\gamma D$$

H 代表水平分量、V 代表垂直分量、D 代表对角分量。E 代表一新的分量，融合了水平、垂直、对角方向的信息。α、β、γ 分别为各分量的权值。实验中分别取为 0.5、0.4、0.1。

E 分量如图 3 所示。

图 3 E 分量

Fig. 3 E component

2.1.2　强度检测

从图 3 可看出,从小波分解后的子图中仍包含噪声,要准确地提取出文本,还需进一步的处理。分析文本区域发现,文本高频分量点应具有一定的强度和密度。

强度是指其高频分量的值不能太低,因此,设置一门限,将低于此值的点滤除。经强度检测后二值化的结果如图 4 所示。

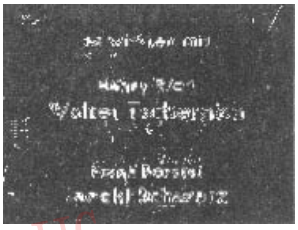

图 4　强度检测后的高频分量

Fig. 4　The high frequency component by intensity detection

2.1.3　密度检测

密度是指单位面积内高频分量点(也可看成边界点)的个数。定义一个大小为 $m \times n$ 的观察窗,用此观察窗分别在水平和垂直两个方向以距离为 $m/2$(水平方向)、$n/2$(垂直方向)扫描图像。分析观察窗里高频分量点的分布。发现:局部窗口边界点的直方图基本成双峰形状,在双峰之间的谷点即为阈值点。去除低于此阈值的点,即可简化复杂背景。经密度检测后二值化的结果如 5 图所示。

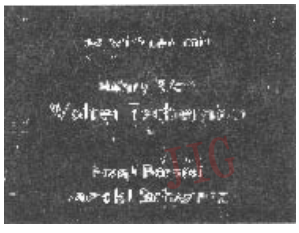

图 5　密度检测后的高频分量

Fig. 5　The high frequency component by density detection

2.1.4　基于 LH 的凸台检测

传统的数字图像灰度直方图(gray histogram,即 GH)是灰度级的函数,它反映出图像中各个不同灰度级的像素数,其中像素的统计是在整幅图像中进行的。因此,它不具有描述图像局部特征的能力。定义一种局部直方图(local histogram,即 LH)如下:

给定一图像 $f(x,y)$,对于其中由若干连续的行或列(数量为 L)所构成的任意子图像,可分别按行或按列生成局部灰度直方图 LH。即

$$LH(k) = N_k / N \times 100\% \quad k = 1, \cdots, L$$

其中,k 为子图像的像素行(列)号,N 表示一行(列)中的像素总数,N_k 是子图像第 k 行(列)中的灰度或高频分量点的像素数。

LH 与传统的 GH 有相似之处,但又有不同,它反映图像的某些局部特征,具有一定的实用价值。

图像中某些局部区域,如文本,其灰度或高频分量点与其邻近的背景区域形成较大反差。因此,这些局部区域使 LH 函数值发生突变,在 LH 图中生成凹谷(或凸台)。

定义　在 LH 中,如果存在一个具有一定宽度的连续区域,其函数值明显高于相邻区域,且边界比较陡峭,则称之为凸台。

从定义中知道凸台有如下的特点:边界陡峭,有一段较大高度的连续区域,两边函数值很小。利用这个特点,实现了相应的凸台检测算法。

凸台检测算法:

(1)输入行/列的 LH 值。

(2)逐段计算 LH 值大于一阈值的凸台长度。将凸台的起始位置和长度加入到二元数组 S 中。

(3)计算 S 中相邻凸台之间的距离,若距离低于一阈值,则将各凸台融合;否则,保留原凸台的位置和长度。

(4)输出各凸台的位置和长度。

图 6 是图 5 水平方向的 LH 图。从图中可见,水平方向有几个明显的凸台。设置参数 convex_Height_h、convex_Len_h 分别表示水平方向凸台高度门限和凸台长度门限,只有超过一定高度和长度的凸台才对其进行标记;此外,设置 convex_Dis_h 参数,将水平方向小于一定距离的凸台融合为一个凸台。标记出来的凸台位置即为文本行所在的位置。

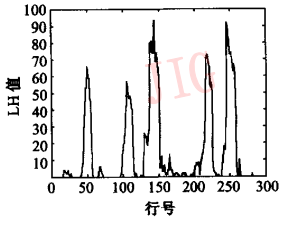

图 6　水平方向 LH 图

Fig. 6　The horizontal LH figure

对应每个文本行所夹的子图像，计算其垂直方向的 LH 值，图 7 为图 6 标记后各文本行在垂直方向的 LH 值。从图中可知，垂直方向 LH 值中由连续的凸台所构成的凸台包，凸台之间的间隙对应字符之间的间隙。参数 convex_Height_v、convex_Len_v 分别表示垂直方向凸台高度门限、凸台长度门限，对具有一定长度和高度的凸台标记其所在的位置。convex_Dis_v 表示融合距离参数，将垂直方向小于一定距离的凸台融合起来。标记出来的凸台位置即为文本列所在的位置。图版 I 图 3 为图版 I 图 1 的文本区域检测结果。

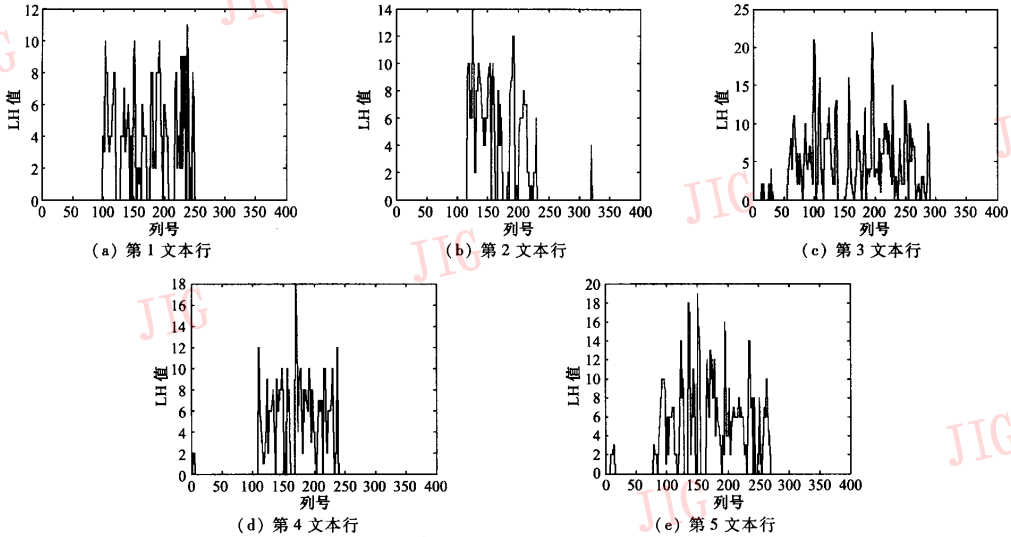

图 7 各文本行所夹子图像垂直方向的 LH 值图(从上往下方向)

Fig. 7 The vertical LH figure of the sub-images between text lines (from up to down)

2.2 文本运动跟踪

由于视频文本一般出现在连续数帧中，用上面的方法检测到文本区域后，用运动跟踪的方法来定位后续帧的文本区域，因此，需对文本区域进行运动估计。而块匹配是运动估计中常用的方法。这种方法是根据一定的匹配准则，在后续帧的搜索区域寻找与当前帧最为接近的那一块，这两个块相对位移的值，即为运动矢量。常用的算法有基于全搜索、三步搜索、菱形搜索等。由于三步搜索、菱形搜索等快速算法容易陷入局部最优解，因此考虑用性能良好的全局搜索方法，由于只对当前帧的文本区域块在后续帧搜索相应的位置，计算量并不大。匹配准则采用最小均方误差，均方误差的表达式为

$$MSE(I_i,I_j)=\frac{1}{w\times h}\sqrt{\sum_{x=0}^{w}\sum_{y=0}^{h}(I_i(x,y)-I_j(x,y))^2}$$

其中，I_i、I_j 分别表示当前文本块和相对应块的亮度值，w、h 表示文本块的大小。

如图 8 所示，最左边一幅图是图版 I 图 3 中一

图 8 跟踪结果

Fig. 8 The tracking result

个文本块的定位结果,右边两幅图是在其他帧对该文本块跟踪的结果。

2.3　文本增强

用上述方法对视频帧文本的定位结果还不能直接送入 OCR 进行识别。这是因为,文本字符的分辨率很低;文本仍然与复杂背景混合在一起(如图 9 所示)。为此,需对文本字符进行增强处理,将文本字符从复杂背景中分离出来。

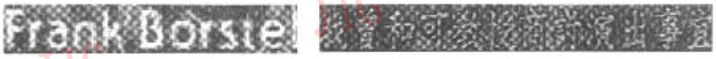

图 9　单帧文本块

Fig. 9　The text block of single frame

观察视频文本发现:为便于阅读,视频文本一般出现在视频连续数帧之中,而且相对文本像素,视频背景像素点的波动明显高于文本像素的波动。为此,将跟踪到的多帧文本区域进行平均来增强文本区域。设 C 为包含文本字幕的视频序列,帧 $f \in C$ 包含文本区域 $\gamma(f)$。帧平均方法通过下式来增强文本区域:

$$\bar{\gamma} = \frac{1}{|C|}\sum_{f \in C}\gamma(f)$$

其中,$|C|$ 代表 C 内所取的帧数。如图 10 是取相隔 5 帧,帧数为 10 帧平均后所得结果。从图中可知,采用帧平均后,背景的复杂性降低,文本区域得到加强。

图 10　多帧平均结果

Fig. 10　The averaging result of multi-frames

2.4　字符二值化

由于取文本区域进行处理,很大部分像素点是文本像素点。首先判断文本是正相还是反相,所谓正相文本即文本亮度高于背景亮度;而反相文本,则相反。为此,先将所检测到的文本区域向 4 个方向扩大两个像素的位置,然后将文本区域上下边界 2 行像素的均值与边界以内中心区域的像素点的均值进行比较,若前者低于后者,则为正相文本;否则,为反相文本。然后,按下式计算阈值:

$$T = m + k \times v$$

其中,m 代表文本区域的均值;v 代表文本区域方差;k 为用户定义的系数。

若为正相文本,则取大于 T 的像素为文本像素,其余为背景像素;若为反相文本,则取小于 T 的像素为文本像素,其余为背景像素。如图 11 为图 10 的二值化结果。

图 11　二值化结果

Fig. 11　The binarization result

从图 11 可以看到,二值化结果中仍存在大量的噪声,通过分析连通分量的方法来滤除噪声点。采用 8-连通分析,对图 10 连通区域进行标记,去除过分小的区域(实验中将 3 个像素以下的区域去掉)将孤立噪声点去除。这样处理后,仍有大片被判为文本的背景区域存在(如图 11“商讨”两个字周围)。幸运的是,为便于阅读,一般对比度很低的文本字符都有轮廓线包围,而有大片被判为文本的背景区域一般与边界相连(因已将文本区域扩大了,文本一般与边界不连),因此,去除与边界相连的区域,获得如图 12 所示的结果。

2.5　OCR 识别结果

将文本和背景黑白反转后(如图 13 所示)送到 OCR 软件进行识别。图 14 是用软件清华紫光 OCR

图 12　连通分量处理结果

Fig. 12　The procession result of connected component

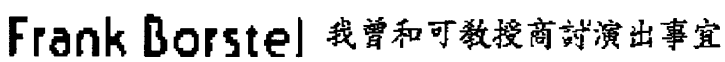

图 13 黑白反转后结果

Fig. 13 The result of inverting black and white pixels

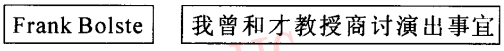

图 14 OCR 识别结果

Fig. 14 The recognization result by OCR

V7.5 识别后的结果。

3 实验结果

对一系列视频片段进行实验,所选取的测试视频帧在不同背景(简单背景或存在许多"似文本"区域的复杂背景)、单行或多行文本、不同语种(中文或西文)、不同字体颜色、不同字体大小,视频序列有静止背景运动文本、运动背景静止文本、运动背景运动文本等多种情况下进行测试。表 1 列出了用该方法检测到的结果,其中引起部分误检有两种情况:即非文本检测为文本或文本检测为非文本。非文本检测为文本的原因主要是复杂背景下存在的与文本相似性很高的区域,而文本检测为非文本的主要原因是由于该文本区域与周围背景对比度太低几乎完全溶入到背景中时的情况。图版 I 图 4 为其他部分视频帧文本区域检测到的结果,图 15 为运动文本运动背景其中一行文本的跟踪结果。采用软件清华紫光 OCR V7.5 对二值化后的字符的最终识别率为 74.5%。从实验结果来看,本文所提出的方法能较好地提取和跟踪视频帧的文本,对文本区域的定位精度较高,识别效果较好。

表 1 文本检测结果

Tab. 1 The text detection result

总帧数	文本区域数	正确检测	误检
265	336	331(98.5%)	54

图 15 运动文本运动背景其中一行文本的跟踪结果

Fig. 15 The tracking result of one line in frames in which text and background are moving

4 结 论

本文用基于小波变换和 LH 的方法检测视频帧文本所在的位置,将检测到的文本区域在后继帧进行跟踪,经过多帧平均、连通分量分析、二值化等步骤将检测到的文本字符送入 OCR 软件进行识别。该方法简单,能快速地定位和跟踪文本区域,且不受文本颜色、字体大小、语种(比如中文或西文)等的限制,定位精度较高,识别效果较好。

参考文献(References)

1 Li Huiping, Doermann David, Omid Kia. Automatic text detection and tracking in digital video [J]. IEEE Transactions on Image Processing, 2000, 9(1):147 ~ 156.

2 Zhong Yu, Zhang Hong-Jiang, Jain A K. Automatic caption extraction of digital videos [A]. In: Proceedings of International Conference on Image Processing (ICIP '99) [C], Kobe, 1999: 24 ~ 27.

3 Sato Toshio, Kanade Takeo, Ellen K Hughes, *et al*. Video OCR for digital news archives [A]. In: IEEE Workshop on Content-Based Access of Image and Video Databases(CAIVD '98) [C], Bombay, India, 1998:1405 ~ 1413.

4 Kim Hae Kwang. Efficient automatic text location method and content-based indexing and structuring of video database video [J]. Journal of Visual Communication and Image Representation, 1996, 7(4):336 ~ 344.

5 Jain Anil K, Bhattacharjee Sushil. Text segmentation using Gabor filters for automatic document processing video [J]. Machine Vision and Applications, 1992, 3(5):169 ~ 184.