第 34 卷 第 3 期 计 算 机 工 程 2008 年 2 月
Vol.34 No.3 Computer Engineering February 2008
·博士论文· 文章编号：1000—3428(2008)03—0015—03 文献标识码：A 中图分类号：TP391

基于点模式匹配的视频文字跟踪和笔画提取

马 瑞，王家廞
(清华大学计算机系智能技术与系统国家重点实验室，北京 100084)

摘 要：给出一种在复杂背景下的视频文字跟踪和文字笔画提取方法。用基于 Harris 角点特征的点模式匹配法跟踪视频序列中静止和运动的文字，以确定文字序列的时间属性，比较了采用图像整体像素匹配和点模式匹配的跟踪精度。用基于多帧融合思想的前景/背景识别算法提取视频文字笔画并作 OCR 识别。实验结果显示，点模式匹配的跟踪算法比图像整体像素匹配的算法跟踪精度更高，在图像背景复杂、变化快的情况下，基于多帧融合的文字笔画提取方法优于传统的二值化方法。

关键词：视频文字跟踪；点模式匹配；文字笔画提取

Video Text Tracking and Stroke Extraction Based on Corner Feature Matching

MA Rui, WANG Jia-xin
(State Key Laboratory of Intelligent Technology and Systems, Dept. of Computer Science, Tsinghua University, Beijing 100084)

【Abstract】This paper proposes a video text tracking and text stroke extraction method under complex background. A point matching method based on Harris corner features is introduced to track text objects. The performance is evaluated by the comparison with SSD-based tracking method. A multi-frame-based foreground/background recognition algorithm is proposed to extract text strokes for optical character recognition. The efficiency and robustness of the point matching method for video text tracking and the text stroke extraction algorithm are proved by objective and thorough experiments on TV serials and movies.

【Key words】 video text tracking; corner feature matching; text stroke extraction

1 概述

在媒体信息处理领域，相比其他视觉信息，视频流中的文字信息能提供更加可靠、详实的语义信息，在基于内容的媒体挖掘和媒体检索中占有重要地位，是这一领域的研究热点。视频中的文字通常有 2 种，即存在于景物当中并被拍摄下来的景物文字，以及各种由机器生成并嵌入图像画面的叠加文字。叠加文字含有关于视频内容的大量信息，是研究的重点，本文的目的是要实现一个快速、稳健的视频文字检测和跟踪算法，并将文字笔画提取出来，作为各种文字识别算法的输入。

已有的视频文字检测和提取算法多基于静态单幅图像的处理[1-2]。不同于静态图像文字，视频文字具有冗余性，出现在连续的若干帧中。已有的算法提出用多帧融合的思想[3-4]增强文字区域，因为文字颜色沿时间比较稳定，采用多帧信息有助于识别变化的背景像素，但该方法需要先获得配准的文字图像序列，也就是文字跟踪问题。文献[2]提出用图像块的文字像素颜色差值作为匹配度，在局部区域内搜索一个最佳的匹配位置。该方法需要根据估计的文字颜色识别出文字像素，而复杂背景下的文字颜色常常难以估计，且当背景与文字颜色相近时，无法正确识别文字像素。文献[5]提出基于边缘图的匹配，但仅限于静态视频文字的跟踪。文献[4]提出一种跟踪运动文字的算法，采用基于SSD的跟踪方法，能够解决一般性的仿射变换下的跟踪问题。

2 视频文字检测

视频文字检测的目的是通过扫描视频图像序列，在单帧图像上检测文字图像区域。为了利于辨认，文字区域通常较图像的其他部分具有更加丰富的角点和边缘特征，与背景颜色有较大的对比度。已有的算法多采用抽边、离散余弦变换或小波变换的特征，提取图像中高频信息丰富的区域，作为文字的候选区域，再通过一些先验知识，如高度、比例等约束和修正。

本文检测视频文字主要思路是：扫描视频流，在单帧图像上检测 Harris 角点特征，利用图像区域的角点密度信息标示出文字块和非文字块。

Harris算子是C.Harris和M.J.Stephens提出的一种角点特征提取算子[6]。其计算简单，能均匀合理地提取角点特征，在特征提取、目标识别和配准等方面有广泛应用。其计算方法如下：首先计算图像x和y方向的梯度D_x和D_y，对每一个像素点p，定义其局部结构矩阵为p点周围$d\times d$邻域内x和y方向的梯度和，公式为

$$\boldsymbol{c}=\begin{bmatrix}\sum D_x^2 & \sum D_xD_y\\ \sum D_xD_y & \sum D_y^2\end{bmatrix} \tag{1}$$

求解这个矩阵的特征值，即求解$\det|\boldsymbol{c}-\lambda\boldsymbol{I}|=0$，其中，$\lambda=[\lambda_1,\lambda_2]$是特征值向量；$\boldsymbol{I}$是单位矩阵。如果较小的特征值大于一个预先设定的阈值ε，那么p点就检测到一个Harris角点。直观上理解，角点特征就是在图像邻域内 2 个正交的方向上都有较大的梯度。用 10×10 像素的窗口滑动过整个图像，滑动步长

作者简介：马 瑞(1979－)，女，博士研究生，主研方向：模式识别，图像处理；王家廞，博士生导师

收稿日期：2007-02-20 **E-mail：**reese.marlin@gmail.com

在x和y方向均为 5 个像素，计算窗口内图像的角点密度，根据文字区域角点密度较大的性质，将窗口标示为文字块和非文字块。通过求文字连通域的最小包围矩形确定文字区域。

3 基于角点模式匹配的文字跟踪

本节的目的是在视频流中对检测到的视频文字进行精确跟踪，确定其起始和结束帧号以及在每一帧中的位置和尺度。基本的思想是取出当前视频帧中检测到的一句文字图像块，提取待匹配的特征，并在相邻的视频帧中搜索和匹配。

在视频文字的跟踪算法中，最常用的方法是利用图像的像素信息和最小化像素均方误差和 (Sum of Squared Distance, SSD)的方法，但该方法在背景变化强烈时容易失效[2,4]。文献[5]中提出用Canny边缘位图做匹配，如果当前帧相应区域的边缘位图与文字对象的边缘位图的相似度大于某个阈值时，则认为匹配成功。这里仍然利用Harris角点集合作为待匹配的特征，由于背景产生的角点远少于背景边缘的比例，而相同文字内容的图像产生的角点特征集合是相当稳定的，其特征点集合中角点的个数和位置变化小，因此本文设计了基于角点特征集合的匹配方法，来度量 2 个图像区域的相似程度，达到文字跟踪的目的。

首先采用点模式匹配在视频流中跟踪静态的文字序列，确定文字图像序列的结束/起始帧，只要在相邻帧中的对应位置进行特征匹配，无须搜索。

对于第 n 帧检测到的一个文字对象，其图像区域由矩形$R=(x,y,w,h)$标示，描述了图像区域的中心点坐标和宽、高，首先在R内计算参考帧的角点特征集合$T=\{p_i\}$, $i=1,2,\cdots,N_T$，N_T是参考帧角点的个数。在第 $n+1$ 帧计算矩形区域R内的角点特征集合作为候选集$C=\{q_j\}$, $j=1,2,\cdots,N_C$, N_C是候选帧角点的个数。计算角点集合T与C中位置匹配的特征点对的个数N_m。需要说明的是，由于图像噪声、背景变化等因素，2 个点之间的位置匹配允许一个很小的误差范围，该值与上述角点特征检测时d邻域的选择和检测到的文字样本高度有关。匹配的特征点数按下式计算：

$$N_m=(N_m(T,\ C)\ +\ N_m(C,\ T))/2 \tag{2}$$

其中，

$$N_M(T,C)=\sum_{p_i\in T} I\left(\min_{q_j\in C}\|p_i-q_j\|\right) \tag{3}$$

$$N_M(C,T)=\sum_{q_j\in C} I\left(\min_{p_i\in T}\|q_j-p_i\|\right) \tag{4}$$

其中，$I(x)=\begin{cases}1 & x\le d\\ 0 & \text{otherwise}\end{cases}$；$\|\bullet\|$定义了两点之间的距离范式(欧式距离)。式(3)解释如下：对每一个 $p_i\in T$，计算它到集合C的最小距离，如果该距离不大于一个很小的允许误差范围，则存在一对匹配特征点，统计所有集合T到集合C的特征点匹配个数$N_M(T,\ C)$。$N_M(T,\ C)$是集合C到集合T的特征点匹配个数。那么角点模式集合T与C之间的匹配度可计算为

$$M(T,\ C)\ =\ N_M/\max(N_T,\ N_C) \tag{5}$$

由于N_M总是不大于N_T和N_C，匹配值界于 0 到 1 之间。N_M越大表示 2 个角点集合的匹配程度越高。

基于这样的度量方法，可以通过设定阈值来获得匹配结果。阈值由实验确定，基于 2 组随机选择的样本。一组样本中有 181 对包含文字的图像，每一对中的文字内容一样，但取自于不同的视频帧。另一组样本中有 560 对文字图像，每一对中的文字内容都不一样。计算两组样本中每一对图像的角点特征集合匹配值，其分布如图 1。可以看到，相同文字内容的图像角点集匹配值分布多集中于 0.6~0.9，而不同文字内容的图像角点集匹配值分布多位于 0.4 以下。实验中，在不同的视频流上选取阈值 0.5，采用中文和英文字幕测试，可以达到平均 1 帧以内的跟踪误差，而基于图像整体像素匹配的平均跟踪误差在 10 帧以内。

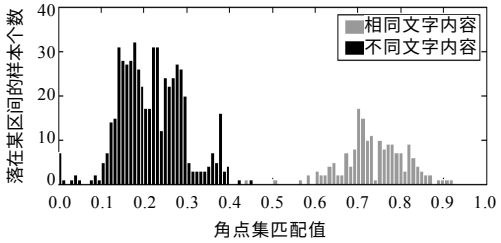


图 1　图像角点模式匹配值分布

对于视频中运动的文字，仍采用角点模式做匹配特征，但由于文字位置不确定，因此需要计算文字区域在水平和垂直方向上的最佳位移量 $(u^{\text{opt}}, v^{\text{opt}})$。仍如上定义，在参考帧中待匹配文字区域$R$内提取角点模式集合 T 。由于一般视频中文字在平面内做旋转和尺寸缩放的并不多见，因此本文只考虑纯位移的情况。在搜索中，定义待匹配图像区域由R进行位移变换确定，即$R'=(x+u,y+v,w,h)$，u, v分别为x, y方向上的位移。由矩形R'确定的图像区域中的角点模式集合为C。 最佳的位移参数$(u^{\text{opt}},\ v^{\text{opt}})$可通过优化角点模式集合$C$与$T$的匹配度获得，即

$$(u^{\text{opt}},v^{\text{opt}})=\arg\max_{|u|\le dx,|v|\le dy} M(T,C) \tag{6}$$

其中，$(dx,\ dy)$是搜索区域，$M(T,\ C)$按式(2)计算。通过优化式(6)即可确定最大的匹配度以及$(u^{\text{opt}}, v^{\text{opt}})$，如果该最大匹配值大于文字匹配所需的最小相似性约束，本文选择 0.5，则在相邻帧中跟踪到匹配的视频文字，否则，就没有跟踪到。搜索区域大小的选择与事先假设的文字运动速度有关。

4 文字提取

一般的文字识别软件要求输入图像中背景简单干净，文字笔画清晰连贯。由于视频文字的背景颜色复杂多变得不到较好的识别效果，通常文字笔画相比于背景具有较大的亮度，因此以往的算法多采用二值化的方法在静态单帧图像上提取文字笔画[1-2]。而在视频流中，可以利用多帧融合技术进行文字区域的增强[3]，提高文字分割的精度。本文的算法也是利用多帧的冗余信息来提取文字笔画，通过以下 2 个步骤识别背景和文字像素：

(1)采用多帧融合思想识别候选文字笔画像素集合O_c。首先利用角点匹配算法获得文字语句的开始帧号i和结束帧号$i+n$，然后采用最大最小帧搜索[3]方法获得最大$I_{\max}$和最小$I_{\min}$图像：

$$I_{\max}(x,y)=\max(I_i(x,y),I_{i+1}(x,y),\cdots,I_{i+n}(x,y))$$

$$I_{\min}(x,y)=\min(I_i(x,y),I_{i+1}(x,y),\cdots,I_{i+n}(x,y))$$

不同于文献[3]中只采用最小图的增强做法，本文考察$\Delta I\ =|I_{\max}-I_{\min}|$中像素的变化。由于文字像素值是稳定的，如图 2(a)，在复杂多变的背景情况下，背景像素的ΔI值比文字像素的大，因此候选文字笔画像素集合O_c定义为

$$O_c=\{p=(x,y)\mid \Delta I(p)<\sigma_o\}$$

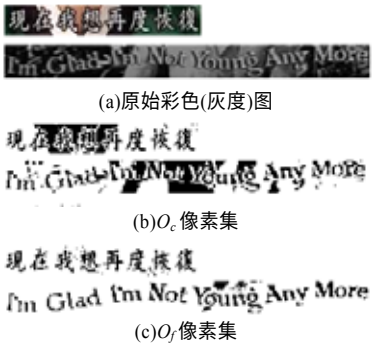
其中，经验阈值σ_o由实验确定；O_c像素集如图 2(b)所示。

(2)利用前后帧信息识别最终文字笔画像素集合O_f。通常在背景变化的情况下经过上一步的识别，能够得到较好的文字笔画提取效果，但有时文字的背景是静止的，这样会保留

大量的背景像素，这一步利用同一个镜头序列中，文字序列的前帧i-1 和后帧$i+n$+1 识别O_c中的文字像素，这 2 帧图像是不含有文字语句的。在同一个镜头序列中，相邻 2 帧图像背景变化小，而文字的有无会产生较显著的像素值变化，因此，计算文字切换处两帧差$\Delta I_f = |I_i - I_{i-1}|$和$\Delta I_b = |I_{i+n} - I_{i+n+1}|$，定义如下：

$$O_f = \{p = (x,y) \mid \Delta I_f(p) > \sigma_f, \Delta I_b(p) > \sigma_b\}$$

其中，σ_f, σ_b为实验确定的阈值；从O_c中识别出最终的文字像素集合O_f如图 2(c)所示。



(a)原始彩色(灰度)图

(b)O_c像素集

(c)O_f像素集

图 2　视频文字笔画提取示例

5　实验结果

本文在不同电影和电视连续剧数据上测试视频文字的匹配算法，测试数据包括中英文文字、简单和复杂背景。实验比较了基于图像块匹配和角点模式匹配方法对文字跟踪的精度，定义序列 i 的起始/结束帧跟踪结果与手工标注结果的差分别为$|\Delta n_{i,\text{begin}}|$和$|\Delta n_{i,\text{end}}|$，视频中的总文字语句数为$n_{\text{total}}$，则平均跟踪误差定义为

$$err_{\text{tracking}} = \frac{\sum_i (|\Delta n_{i,\text{begin}}| + |\Delta n_{i,\text{end}}|)}{n_{\text{total}} \times 2}$$

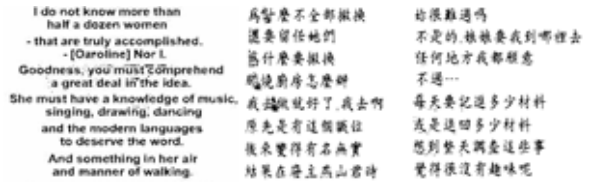
测试的结果如表 1 所示，以角点模式匹配的方法平均跟踪误差在 1 帧以内。部分实验比较及结果见图 3、图 4(测试内容选自《傲慢与偏见》和《大长今》第 31 集)。

表 1　图像 SSD 匹配和角点集合匹配静态视频文字的实验结果

视频序列	语言	文字句数	平均跟踪误差/帧	
			图像SSD匹配[4]	角点模式匹配
1#	英文	358	8.76	0.40
2#	中文	284	10.27	0.25
3#	中文	179	5.50	0.00
4#	中文	25	5.00	0.00
5#	中文	21	6.30	1.26
6#	中文	20	12.20	0.00

(a)原始灰度图　(b)二值化方法　(c)本文的方法

图 3　文字提取笔画方法比较



(a)英文　(b)中文

图 4　视频文字提取结果

6　结束语

基于内容的媒体挖掘、视频理解和检索是近年来研究的热点，视频文字的提取和处理是其中的关键技术。针对视频中常见的复杂背景和未知文字颜色的问题，本文提出了基于角点模式匹配的视频文字跟踪和基于多帧融合技术的文字提取方法，并在随机抽取的视频数据上得到了有效验证。下一步的工作是将提取的文字进行 OCR 识别，更好地进行语义信息的抽取和视频理解。

参考文献

[1] Lyu M R, Song Jiqiang, Cai Min. A Comprehensive Method for Multilingual Video Text Detection, Localization, and Extraction[J]. IEEE Trans. on Circuits and Systems for Video Technology, 2005, 15(2): 243-255.

[2] Lienhart R, Wernicke A. Localizing and Segmenting Text in Images and Videos[J]. IEEE Trans. on Circuits and Systems for Video Technology, 2002, 12(4): 256-268.

[3] Sato T, Kanade T, Hughes E K, et al. Video OCR for Digital News Archive[C]//Proc. of IEEE Workshop Content-based Access Image Video Database. Bombay, India: [s. n.], 1998.

[4] Li H, Doermann D, Kia O. Automatic Text Detection and Tracking in Digital Video[J]. IEEE Trans. on Image Processing, 2000, 9(1): 147-148.

[5] 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[J]. 计算机研究与发展, 2006, 43(9): 1523-1529.

[6] Harris C, Stephens M J. A Combined Corner and Edge Detector[C]// Proceedings of the 4th Alvey Vision Conference. Plessey, United Kingdom: [s. n.], 1988.

（上接第 14 页）

[4] Niculescu D, Nath B. Ad Hoc Positioning System (APS) Using AoA[C]//Proceedings of the IEEE INFOCOM'03. [S. l.]: IEEE Press, 2003: 1734-1743.

[5] Girod L, Bychovskiy V, Elson J. Locating Tiny Sensors in Time and Space: A Case Study[C]//Proceedings of the IEEE International Conference on Computer Design. Freiburg: [s. n.], 2002: 214-219.

[6] Bulusu N, Heidemann J, Estrin D. GPS-less Low Cost Outdoor Localization for Very Small Devices[J]. IEEE Personal Communications Magazine, 2000, 7(5): 28-34.

[7] Nicolescu D, Nath B. Ad Hoc Positioning Systems (APS)[C]// Proceedings of IEEE GLOBECOM'01. [S. l.]: IEEE Press, 2001: 2926-2931.