一种基于聚类的视频字幕提取方法

宋砚[1,2,3]，颜成钢[4,1]，刘安安[1,2]，庞琳[1,2,3]，张勇东[1,2]，唐胜[1,2] 林守勋[1,2]

（1．中国科学院计算技术研究所，虚拟现实技术实验室，北京 100190；

2.中国科学院计算技术研究所 智能信息处理重点实验室，北京 100190；

3.中国科学院研究生院，北京 100049；

4．山东大学威海分校，山东 264209）

摘 要： 本文针对现有方法的不足提出了一种视频中字幕提取的方法。本方法采用多尺度文字定位，并加上文字区域精确化，得到文字区域后运用改进的 K 均值聚类方法对其中的象素点进行聚类，获得二值化的图像，最后用 OCR 软件识别得到文字。本方法应用于网络视频敏感词语检测具有良好的效果，实验证明了该方法的有效性。

关键词： 文字提取；OCR；K 均值聚类

A New Video Text Extraction Method Based on Clustering

Yan Song[1,2,3], Chenggang Yan[4,1], Anan Liu [1,2], Lin Pang[1,2,3], Yongdong Zhang[1,2],

Sheng Tang[1,2], Shouxun Lin[1,2]

(1.Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190) 、

(2.Virtual Reality Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190) 、

(3.Graduate University of the Chinese Academy of Sciences, Beijing 100049)

(4. Shandong University at Weihai, Shandong, 264209)

Abstract: In order to overcome the weakness of existing methods, a new method for video text extraction is proposed. Multi-scale text location and text region refinement is adopted. Then an improved K-means clustering method is used to segment text regions to obtain the binarized image. Finally, OCR software is used to recognize text. The method is used in sensitive words detection in videos from the internet and obtains good results. Experiments have proved the advantage of the proposed method.

key words: Text extraction; OCR; K-means clustering

1．引言

目前，多媒体信息在互联网上以惊人的速度增长，尤其以数字视频最为突出。随着这些视频数据的海量增长，人们在互联网上对于这些信息的浏览需求也日益增加，从而迫切需要对这些数据进行有效的管理、索引和检索。但是目前看来，多媒体内容分析有个难以克服的问题，即低层次的特征不足以描述高层次语义内容。不过，在数字视频中人工添加的字幕则是一种具有高层次语义信息的资源。文字信息相对于视频和音频来说更容易提取，并且可以高度概括相应视频的内容。另外，光学字符识别（Optical Character Recognition）技术已经比较成熟，在很多领域得到广泛的应用。所以，利用视频中的字幕信息来对视频内容进行分析和检索是一种有效的方法。

基金资助：中国国家基础研究项目（973 项目，2007CB311100），中国国家高科技和研究发展项目（863 项目，2007AA01Z416），北京市科技新星项目（2007B071）

作者简介：宋砚（1983-），女，江苏南京，博士在读 Email:songyan@ict.ac.cn

一般来说，现有文字提取方法包括以下几个步骤：文字检测、文字定位、文字增强、文字分割和文字识别。多数方法重点讨论文字定位和文字分割。文字定位常用方法有基于连通分量的方法[1]，基于纹理的方法[2]和基于边缘的方法[3]。基于连通分量的方法可以快速的定位到文字的区域，不过缺点是当背景比较复杂的时候容易失败。基于纹理的方法一般利用机器学习的方法进行分类来判断某一个区域是否属于文字区域。该方法的缺点是需要采集训练样本进行模型的训练。基于边缘的方法对一般文字来说简单有效，不过一些尺寸过大的文字可能存在检

不完整的问题。对于视频文字提取来说，一般的文字分割方法是利用阈值的方法和基于笔画的方法。基于阈值的方法包括 Otsu 方法[4]和 Niblack 方法[5]。这种阈值的方法比较简单，但是对于一些复杂的图像来说效果不理想。基于笔画的方法[6]是运用一些特定的滤波器对文字区域进行选择，各种方向的笔画对各种滤波器响应大于非笔画部分。不过该方法也有一定的局限性，某些笔画相交的部分象素可能会被遗漏。

针对以上问题，我们提出了一种新颖的视频字幕提取方法，并且将该方法运用于网络视频敏感词语的过滤，取得了比较好的结果。该方法包括各尺寸的文字定位，基于聚类的文字分割以及文字识别。文字定位是采用基于边缘的方法，并且针对该方法的缺陷，增加了多尺度变换图像得到图像金字塔来进行定位。文字分割部分主要是利用 K 均值的方法对图像象素点进行聚类，并且针对 K 均值方法的缺点对其进行了改进，自动生成聚类数目和初始中心点。接下来将具体介绍各个步骤。

2．算法介绍

2．1．文字定位

文字定位是为了找到图片或视频帧中出现的各种文字的位置。我们将这一步分为两个步骤：多尺度文字定位和文字区域精确化。基于边缘的定位方法缺点是大字体的文字容易检不完整，所以我们采用了一种多尺度的方法来检测各种字体大小的文字。文字区域精确化将上一步中的得到的错误定位结果去掉，并且进一步精确的定位到文字。

首先介绍多尺度文字定位。我们发现高度小于 6 个象素点的文字很容易被漏检，大于 25 个象素点的文字则容易检不完整。于是我们采用多尺度检测的方法来克服这个问题。将原图缩小到原图的 1/4，即长宽各缩小到原图的 1/2，同时放大到原图的 4 倍，即长宽各放大到原图的 2 倍。这样加上原图组成了一个图像金字塔。在这组图像中实现文字定位方法，各自检测到的文字区域再以“并”的方式融合，即可得到各个尺寸大小文字的区域。

文字定位方法基于边缘，首先提取灰度图像的 Sobel 边缘[7]，然后设定一个阈值将边缘点二值化，留下较强的边缘点。然后建立一个和原图大小同样的标志图。用一个 n*n 的窗口扫描每个边缘图，在我们的试验中 n 取 4。观察每个窗口内的边缘点分布情况，如果在左上、左下、右上和右下四个区域中都有边缘点存在，则将标志图中对应于该窗口的位置设为 1，否则设为 0。于是得到一个二值的标志图。这样，比较大的物体（比如人体）的边界线上的边缘点就在标志图中被除去，而属于文字区域的边缘点则更容易被保留。随后检测标志图中的连通分量，得到每个分量的最大外接矩形。将这些矩形块进行融合即可得到各种尺寸文字的区域。上述过程在图 1 中表示。

文字区域精确化这一步中主要达到以下几个目的：将上一步得到的文字区域紧缩，得到更精确的文字定位；将多行文字分割成单行文字以减少背景；去掉错误定位的区域。我们主要采用文献[8]中的方法。它将文字区域的边缘点向水平方向和垂直方向投影，在投影的结果中定位峰/谷的位置以作为精确化的依据。

(a) (b) (c)

图 1 文字定位的各个步骤结果 (a)二值化的边缘图;(b)标志图;(c)定位的区域

Fig.1 Results for each step of text location (a)binarized edge map;(b) marking map;(c) located areas

2．2．文字分割

文字分割是在文字定位的区域里将文字和背景分割开来，得到二值化的图像，作为接下来文字识别的输入。我们提出了一种基于 K 均值聚类的方法，并且对传统 K 均值做了一些改进。聚类方法在图像分割中应用很广泛，其中 K 均值方法是一种简单有效的聚类方法。不过 K 均值聚类方法的一个缺点就是需要人为指定 K 的值。由于对于不同的图像，类数不是固定的，于是我们采用一种自适应的选择类数和初始聚类中心的方法。

首先介绍一下所用的特征。我们采用彩色图像的 HSV 值，并且将 HSV 空间转换到一个立方体空间：

$$X = V \tag{1}$$

$$Y = (V * S * \sin(H) + 1)/2 \tag{2}$$

$$Z = (V * S * \cos(H) + 1)/2 \tag{3}$$

这里，H∈[0,2π]，S∈[0,1]，V∈[0,1]。我们还增加了一维特征来表示局部边缘点分布。局部边缘点分布特征由下式来表示：

$$f_{LocalEP}(i,j) = \frac{1}{25} \sum_{\substack{i-2<=m<=i+2 \\ j-2<=n<=j+2}} P(m,n) \tag{4}$$

P(m,n)表示边缘点图中位于(m,n)位置点的值。这样一个 4 维的特征空间由 3 维颜色和 1 维局部边缘点分布特征组成。

然后介绍关于自适应选取聚类数和初始聚类中心的方法。将特征空间的每一维都均匀分为 q 个等分，本文中 q 取 8。这样整个特征空间总共有 8*8*8*8 个箱。将这些箱按所含有象素点数目从多到少排序。为了节省运算时间，我们选取前 c 个箱满足以下条件：

$$\sum_{k=1}^{k=c-1} n_k / N < th1 \tag{5}$$

$$\sum_{k=1}^{k=c} n_k / N >= th1 \tag{6}$$

这里，n_k表示第k个箱里的象素点数目，N表示象素点总数目。th1 是一个阈值，一般设在 0.7～0.9 之间。这样可以去掉大量无象素点的箱，以及一些包含象素点数目非常小的箱。每个箱用其中象素点的平均坐标来表示，将第k个箱用一个点表示为o_k(x,y,z,f)。建立一个空的中心表，用来记录被选中的箱。首先，将排完序的箱的第一个（即含有最多象素点的箱）加入中心表。然后剩下的选择标准为：

$$k = \arg \max_k f(bin_k) = \arg \max_k (\alpha d_k + (1-\alpha) n_k / N) \tag{7}$$

$$d_k = \min_{bin_j \in center\ list} \{dist(o_k, o_j)\} \tag{8}$$

这里d_k表示bin_k到中心表中所有箱的最小距离。本文中用的是城市街区距离，并且所有距离都归一化到[0，1]区间。公式 7,8 表示那些含有象素点数目较多并且距离选中的箱较远的箱更容易被选中。参数α代表了这两个因素的重要性，本文中α设定为 0.5。然后这个选取箱的过程迭代进行，每次选出一个箱加入中心表，直到以下两个条件之一满足：没有箱可选；剩下的所有箱的$f(bin_k)$均小于th2。经过实验，th2 取 0.07 时效果最好。

当算法结束时，中心表里的箱的数目就作为类数，中心表里的箱的中心就作为初始聚类中心。然后用传统的 K 均值算法进行聚类，将象素点分为 k 类。实验表明，当 th2=0.07 时，k 的取值范围大多数在 2～4 之间。一般视频中为了让人清楚的阅读字幕，都将字幕设为灰度值最高或者最低。判断文字部分是高亮还是低亮的文献有很多[9]，为了简化问题我们只考虑前者，选择亮度最大的那个类作为文字，剩下的类作为背景，将图像二值化。图 2 表示了分割的结果。(a)是原图，(b)是聚类的结果，这个例子中聚为 3 类，分别用 3 种灰度值表示，(c)二值化的结果。

文字识别是识别输入的二值化图像中的文字的过程。一般这一步都用商业 OCR 软件来完成，我们实验中用的是汉王公司的 OCR 软件 HWOCRSDK1.2。

(a) (b) (c)

图 2 文字分割结果示例

Fig.2 Results of text segmentation

3. 应用

我们将本算法用于网络视频敏感词语过滤，取得了比较好的效果。目前网络上视频数目非常庞大，用户上传的视频占了很大比例，其中有些视频涉及到敏感话题或者不适合未成年人的题材，如果广泛传播会对社会造成不良影响。首先，事先设定好敏感词语的关键词，然后利用我们的算法对网络视频帧中出现的字幕进行分割识别，检测识别出来的文字是否包含这些关键词，如果包含这些词将视频标注出来以做进一步处理。

4. 实验

我们将提出的方法在一组视频上进行测试。测试视频来自 TRCVID 2005 和 2006，包括来自 CCTV，NBC，CNN 和 NTDTV 电视台的节目。在 300 帧中总共有 1343 个中文字符和 5011 个英文字母。帧画面的分辨率是 352*240。

图 3 显示了文字定位的实验结果示例。图中白色框内的部分就是定位的文字区域。可见本算法针对不同情况，不同语言类型，不同文字尺寸和风格都很鲁棒。对于文字定位的评测我们采用文献[10]中的方法：如果一个定位的文字块和一个真实的文字块的重叠面积大于前者的75%以及后者的90%，则认为该定位结果是正确的。然后我们采取的评定文字定位算法的标准如下：

$$\text{查全率} = \frac{\text{正确 LB的数目}}{\text{所有GB的数目}} * 100\% \quad (9)$$

$$\text{错检率} = \frac{\text{错误LB的数目}}{\text{所有LB的数目}} * 100\% \quad (10)$$

其中，LB 表示定位的文字块，GB 表示真实的文字块。

我们在表 1 中列出了实验的结果。可见，第一步是为了尽量检测到所有的文字，所以不关注错检率。这样，第一步检测的结果查全率很高并且错检率也很高。但是错检率可以通过第二步的文字区域精确化降低。可以看出错检率从 21.59%降低到了 1.94%。

（a） (b)

图 3 文字定位结果示例

Fig.3 Results of text location

表 1 文字定位实验结果

Tab.1 Experimental results of text location

	查全率	错检率
多尺度文字定位	98.38%	21.59%
文字区域精确化	91.65%	1.94%

在文字分割这一步，实验结果用 OCR 的识别率和查准率来评测：

$$识别率 = \frac{正确识别字符的数目}{所有真实字符的数目} * 100\% \quad （11）$$

$$查准率 = \frac{正确识别字符的数目}{所有识别字符的数目} * 100\% \quad (12)$$

将我们的算法和其他两个传统方法做了对比。Otsu 的方法在[4]中介绍，Niblack 方法在[5]中介绍。对比结果在表 2 和表 3 中列出。我们的方法对中文和英文分别达到了 81.15%和 92.11%的识别率。对于英文的实验结果来说，我们的方法比另外两种方法好一些，但对于中文结果来说增加了将近 20%的准确率。这是因为我们测试集中的中文视频比英文的更加复杂一些，并且中文 OCR 的识别比英文的对于文字分割结果更加敏感。当情况比较简单的时候阈值法可以获得良好的结果，但是当出现比较复杂的情况时阈值法的结果就不够理想，而这也证明了我们方法的鲁棒性。

表 2 中文识别结果

Tab.2 Experimental results of Chinese recognition

	识别率	查准率
Otsu	56.34%	59.12%
Niblack	64.90%	68.24%
本文方法	81.15%	83.11%

表 3 英文识别结果

Tab.3 Experimental results of English recognition

	识别率	查准率
Otsu	89.12%	90.28%
Niblack	82.71%	86.31%
本文方法	92.11%	93.45%

5. 总结

现有的视频文字提取方法在文字定位和文字分割方面都有一些缺点，本文针对这些缺点提出了一种新颖的文字提取方法。用多尺度的方法检测不同尺寸大小的字体，并且运用文字区域精确化方法进一步提高这一步的准确性，然后将改进的 K 均值算法运用于文字分割步骤，取得了良好的试验效果。

参考文献：

[1] A. K. Jain, B. Yu. Automatic text location in images and video frames. In Pattern Recognition, pp.2055-2076, IEEE Computer Society Press, Aug. 1998.

[2] K. I. Kim, K. Jung, H. Kim. Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. IEEE Transaction on PAMI, 25(12):1631-1639, 2003.

[3] V. Wu, R. Manmatha, E.M. Riseman. Textfinder: an automatic system to detect and recognize text in images. IEEE Transaction on PAMI, 21(11):1224-1229, 1999.

[4] N. Otsu. A Threshold selection method from Gray-Level Histograms. IEEE Transaction on Systems, Man and Cybernetics, 9(1):62-66, 1979.

[5] J. Xi, X.S. Hua, X. R. Chen, et al. A Video Text Detection and Recognition System. In IEEE International Conference on Multimedia Expo, pp. 873-876, IEEE Computer Society Press, Aug. 2001.

[6] T. Sato, T. Kanade, E.Hughes, et al. Video OCR for Digital News Archives. In IEEE Workshop on Content-Based Access of Image and Video Database, pp.52-60, IEEE Computer Society Press, India, 1998.

[7] I. Sobel. An isotropic 3*3 image gradient operator. Machine Vision for Three-Dimensional Scenes, 376-379, 1990.

[8] M. R. Lyu, J. Q. Song, M. Cai. A Comprehensive Method for Multilingual Video Text Detecion, Localization, and Extraction. IEEE Transaction on Circuits and Systems for Video Technology, 15(2): 243-255, 2005.

[9] A. Wernicke, R. Lienhart. On the segmentation of text in videos. In IEEE International Conference on Multimedia Expo, pp. 1511-1514, IEEE Computer Society Press, Aug. 2000.

[10] Q. X. Ye, Q. H. Huang, W. Gao, et al. Fast and robust text detection in images and video frames. Image and Vision Computing, 23(6):565-576, 2005.