利用小波变换和 K 均值聚类实现字幕区域分割

王建宇[1)] 张 峰[1)] 周献中[2)] 史迎春[3)] 骆 文[1)]
[1)](南京理工大学自动化学院 南京 210094)
[2)](南京大学工程管理学院 南京 210093)
[3)](武汉通信指挥学院仿真中心 武汉 430010)
(jianyu_wang2000@yahoo.com.cn)

摘 要 提出一种字幕区域分割算法. 首先对图像做小波变换和重构,并抽取字幕区域特征,再分块计算统计特征;然后对子块进行 K 均值聚类,实现字幕区域分割. 与已有算法相比,该算法简单,不需要设置阈值. 实验结果表明,即使在复杂背景下,对于字体、大小和位置都不确定的字幕,该算法仍具有良好的分割效果.

关键词 字幕分割;小波变换;K 均值聚类
中图法分类号 TP391.4

Segmentation of Caption Region Using Wavelet Transform and K-Mean Clustering

Wang Jianyu[1)] Zhang Feng[1)] Zhou Xianzhong[2)] Shi Yingchun[3)] Luo Wen[1)]
[1)](*School of Automation*, *Nanjing University of Science & Technology*, *Nanjing* 210094)
[2)](*School of Management and Engineering*, *Nanjing University*, *Nanjing* 210093)
[3)](*Simulation Center*, *Wuhan Institute of Communication Command*, *Wuhan* 430010)

Abstract An algorithm is proposed in the paper to segment caption region. By the algorithm, firstly, the caption features are extracted by wavelet transformation and reconstruction on the image, and secondly, statistic features are calculated block by block. At last blocks are classified by K-mean clustering method. In comparison with other algorithms, the algorithm is simpler and no requirement for setting any threshold. Experimental results show that the proposed algorithm performs well, even for captions with unknown font, size, and position.

Key words caption extraction; wavelet transform; K-mean clustering

0 引 言

视频中的文本和字幕,如新闻标题、节目内容、播出时间、工作人员名单、演职员表以及各类比赛的比分等,具有丰富的语义信息. 将视频字幕自动分割并识别出来,对基于内容的视频自动理解、索引和检索非常有意义.

字幕识别的前提是从视频图像中分割出字幕区域,分割结果的好坏直接影响识别效果. 然而,由于实际视频图像的背景往往比较复杂且不可预测,同时字幕的字体、大小、出现位置也不能确定,因此从视频图像中准确定位字幕区域并不是一件容易的事情. 许多研究人员为此做了大量工作,概括起来主要有以下 4 类方法:1)进行边缘检测以获得文本的边缘信息[1-2],并分别在水平和垂直方向投影统计;然后通过多个阈值和边缘尺寸限制来确定字幕区域.这类方法虽能快速检测文字,但需要预先设置多个阈值,因此通用性不好,检测错误率也较高. 2)针对字幕文字通常具有相似颜色这一特点,用图像分

收稿日期:2005-11-03 修回日期:2006-05-24
基金项目:江苏省自然科学基金(BK2004137)

割[3]、颜色聚类[4-5]或连通区域分析[6]等方法把文字从背景中分割出来. 但由于文字颜色不固定,可能存在多种颜色,因此对于背景复杂的图像和视频检测效果不够理想. 3)根据图像纹理特征[3,6-7]判断某一个像素点或像素块是否属于文字. 这类方法能适应复杂背景,但计算量大,十分耗时,且算法鲁棒性不好. 4)将图像切分成若干子块,用事先训练好的学习分类器(如支持向量机[8]、神经网络[9]等)对所有子块进行分类,以得到字幕和非字幕 2 类子块. 这类算法检测准确率较高,但计算复杂,且分类效果受训练样本影响较大.

文字一般采用与背景有着强烈对比度的颜色,因此字幕区域表现出比其他区域更高的空间频率. 小波变换具有多尺度、多分辨率的特性,通过对图像进行小波变换并重构高频细节,可以强化字幕区域的特征. 本文提出了一种基于小波变换的 K 均值聚类字幕区域分割算法. 相对于上述 4 类方法,该算法简单,不需要设置阈值,在复杂多变的背景下,对字体、大小、位置都不确定的字幕,仍能保持较高的正确率.

1　算法简介

本文算法流程如图 1 所示,首先对原图像进行两级小波分解,并分别重构两级分解得到的高频细节 然后将所有重构图像分别切分成大小为 $N\times N$ 的若干子块,计算每一子块的统计特征,形成各子块的特征向量 最后用 K 均值聚类算法对所有子块进行聚类分析,将其划分为字幕和非字幕 2 类,从而完成字幕区域的粗分. 由于某些背景块的统计特征和字幕块的统计特征相似,某些字幕区域内子块的特征又和非字幕块的统计特征相似,因此粗分结果中可能存在噪声块,而得到的字幕区域可能存在空洞. 为了去除噪声块并得到更完整的字幕区域,还需要对粗分结果进行后续处理. 本文采用了连通性分析的方法去除噪声块,并用形态学方法填补空洞,以保持字幕区域的完整性.

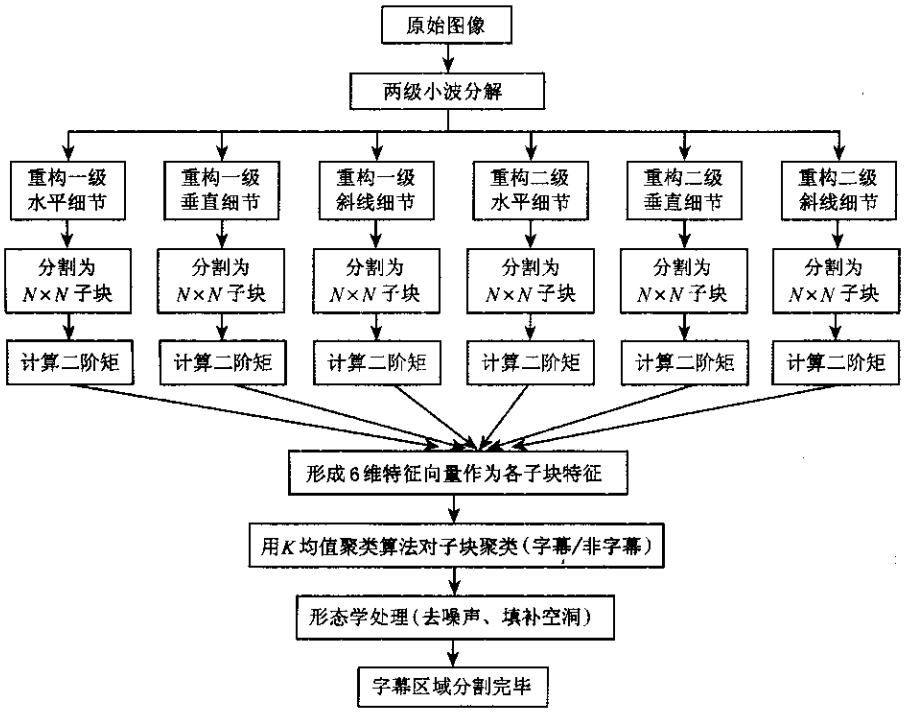


图 1　本文算法流程示意图

2　基于小波变换的特征提取

2.1　小波变换

小波变换多尺度、多分辨率的特性,为其在不同尺度上分析和表征信号提供了一个精确和统一的框架. 图像的小波分解流程如图 2 所示,相应的图像小波分解公式为

$$\boldsymbol{A}_{j+1}^{d}f = \boldsymbol{A}_{j}^{d}f \oplus \boldsymbol{D}_{j}^{HL}f \oplus \boldsymbol{D}_{j}^{LH}f \oplus \boldsymbol{D}_{j}^{HH}f.$$

对图像进行小波分解可将其分为低频部分

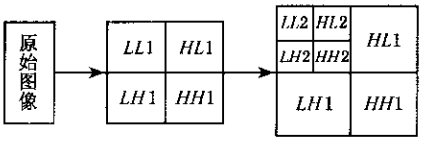


图 2　小波分解流程

$A_{j+1}^{d}f$(近似图像)和高频部分(水平细节 $D_j^{HL}f$、垂直细节 $D_j^{LH}f$、斜线细节 $D_j^{HH}f$). $A_j^{d}f$ 是原图像的近似表达,能够提供丰富的纹理特征;高频部分 $D_j^{HL}f$, $D_j^{LH}f$ 和 $D_j^{HH}f$ 则能够提供图像的边缘信息. 由于字幕区域通常显现出比图像其他区域更高的空间频率,因此小波分解后的细节分量能较好地刻画字幕特征,文献[5,9-10]就是通过提取小波特征作为字幕的特征进行分类的. 图 3 a 所示为一幅含有文本的图像,图 3 b 所示为其一级小波分解图. 从图 3 中可以看出,字幕区域的高频特征在图像细节里得到了充分体现.

a 原始图像 b 一级小波分解结果

图 3 图像一次小波分解图

本文选择 Haar 小波提取字幕特征,因为 Haar 小波在检测边缘信息时具有良好的性能,并且计算简单,可用掩模运算来实现. Haar 小波的尺度函数和小波函数分别为 $f(x)=\sum_{k\in\mathbf{Z}}p_k f(2x-k)=f(2x)+f(2x-1)$ 和 $W_H(x)=\sum_{k\in\mathbf{Z}}q_k f(2x-k)=f(2x)-f(2x-1)$. 其中,$f(x)=\begin{cases}1, & 0\leqslant x<1\\ 0, & \text{其他}\end{cases}$;$p_i=q_i=\begin{cases}1, & i=0,1\\ 0, & i\geqslant 2\end{cases}$.

2.2 特征提取

首先对图像 $I(x,y)$ 进行两级子波分解. 由于低频分量 $A_j^{d}f$ 只是原图像的近似,因此只重构高频细节分量 $D_j^{HL}f$, $D_j^{LH}f$ 和 $D_j^{HH}f$;然后将所有重构的图像分别切分为大小为 $N\times N$ 的子块(本文中 N 取 16),并将所有子块的二阶中心矩(μ_2)作为各子块的特征量,即每个特征向量含有 6 个特征,计算公式为

$$E(\boldsymbol{I})=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\boldsymbol{I}(i,j),$$

$$\mu_2(\boldsymbol{I})=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(\boldsymbol{I}(i,j)-E(\boldsymbol{I}))^2;$$

其中,N 为子块的边长,$\boldsymbol{I}(i,j)$ 为各子块图像数据,$E(\boldsymbol{I})$ 为各子块的均值.

特征提取是在整幅图像上进行的,因此字幕本身的位置并不会对分割结果造成影响. 而选取的特征是每个子块经过小波变换后得到的统计特征,已经去掉了文字所具有的字体和大小信息,所以本文算法可以更好地适应字体和大小的变化.

3 基于 K 均值聚类的字幕区域分割

3.1 K 均值聚类

K 均值聚类是模式识别中的经典算法,它具有计算简单,能够动态聚类,自适应性强等特点,并有着广泛的应用领域,尤其是当解决模式分布呈现类内团聚状的问题时,能取得很好的聚类结果. 该算法的基本思想是取定 K 类,并选取 K 个初始聚类中心,按最小距离原则将各模式分配到 K 类中的某一类,之后不断地计算类心,同时调整各模式的类别,最终使各模式到其判属类别中心的距离平方之和最小. 算法步骤如下:

Step1. 任选 K 个模式特征矢量作为初始聚类中心 $z_1^{(0)}$, $z_2^{(0)},\cdots,z_K^{(0)}$,令 $n=0$.

Step2. 将待分类的模式特征矢量集 $\{x_i\}$ 中的模式逐个按最小距离原则划分给 K 类中的某一类,即如果 $d_{il}^{(n)}=\min_j\left[d_{ij}^{(n)}\right]$, $i=1,2,\cdots,N$,则判定 $x_i\in\omega_l^{(n+1)}$,其中 $d_{ij}^{(n)}$ 表示 x_i 和 ω_j^n 的中心 z_j^n 的距离,n 表示迭代次数. 于是,产生新的聚类 $\omega_j^{(n+1)}(j=1,2,\cdots,c)$.

Step3. 计算重新分类后的各类心

$$z_j^{(n-1)}=\frac{1}{n_j^{(n+1)}}\sum_{x_i\in\omega_j^{[k+1]}}x_i,\ j=1,2,\cdots,K,$$

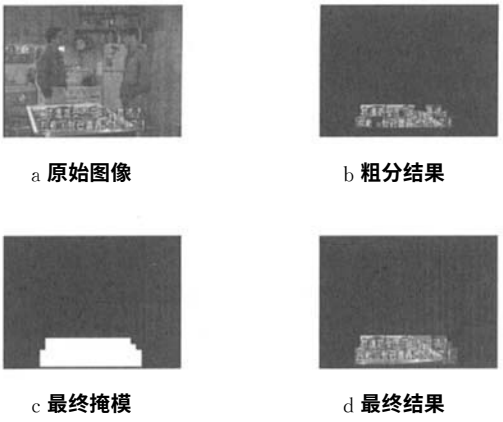
其中 $n_j^{(n+1)}$ 为 $\omega_j^{(n+1)}$ 类中所含模式的个数.

Step4. 如果 $z_j^{(n+1)}=z_j^{(n)}(j=1,2,\cdots,K)$,则结束算法;否则 $n=n+1$,转 Step2.

3.2 字幕区域分割

从本质上讲,字幕区域检测就是 2 类模式分类问题,即将图像中所有子块分为字幕与非字幕 2 类. 此前有研究人员使用支持向量机或神经网络的方法对其分类,但是这 2 种方法都比较复杂,且分类效果受训练样本影响较大. 因此,本文提出使用 K 均值聚类算法对子块进行聚类分析,所有的子块都被划分为 2 类——字幕或非字幕;然后将这 2 类的聚类中心与事先从多个样本得到的聚类中心相比较,以确定哪一个聚类中心代表字幕,哪一个代表非字幕;此时得到一个如图 4 b 所示的字幕区域粗分结果. 由于 2 个聚类中心的相差比较大,因此样本的选择对于类别的判定影响是很小的. 如果得到的 2 个聚类中心和已知非字幕聚类中心的距离都小于已知字

幕聚类中心,说明当前图像中没有字幕出现,因此本文算法也能够检测一幅图像中是否存在字幕.



a 原始图像　　b 粗分结果

c 最终掩模　　d 最终结果

图 4　字幕分割结果

在背景图像十分复杂的情况下,一些表现和字幕块类似特性的背景图像被错判为字幕块是难以避免的. 所幸的是这种情况通常表现为一些小的孤立区域,而视频、图像中的文字大多具有水平聚集排列或垂直聚集排列的特点,因此通过连通性分析就能消除绝大多数孤立噪声块. 此外,字间空格或标点符号等由于在块中比例过小,很容易被聚类到非字幕块中,形成如图 4 b 所示字幕区域的空洞. 另外,以固定尺寸划分图像子块有可能造成边界文字的缺失,即文字的一部分划分到字幕块,另一部分划分到非字幕块. 针对区域空洞和文字缺失的问题,本文分别采用形态学的方法加以处理. 其具体算法如下:

Step1. 根据聚类结果标注所有初选字幕块,1 表示字幕块,0 表示非字幕块,每个块对应一个“像素”,于是所有子块形成一幅二值图像 $I_{块}$,接着对 $I_{块}$ 进行连通性分析,计算所有连通区域面积,面积小于 N(N 为限定的字幕区域至少覆盖的块数,本文选 $N=3$)的连通区域包含的初选字幕块被判断为噪声块,并从 $I_{块}$ 中去除.

Step2. 去除噪声块后,得到的字幕区域内可能还存在空洞,选择适当的结构算子(本文选取 3×3 的结构算子 E_1)

$$E_1=\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$$

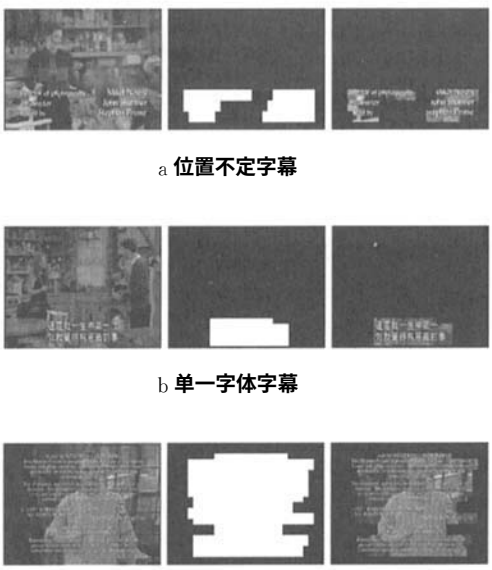
对 $I_{块}$ 进行膨胀处理,可以填补空洞;再用同一个结构算子对 $I_{块}$ 进行腐蚀处理,以保证字幕区域大小在处理前后一致.

Step3. 将字幕子块对应的所有像素标记为 1,其余标记为 0,可得到字幕区域掩模图像,如图 4 c 所示.

Step4. 保持边界文字的完整性. 选择适当的结构算子(本文选取 2×2 的结构算子 E_2)$E_2=\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}$ 对掩模图像进行膨胀运算;将结果与原始图像进行掩模运算,得到最终字幕区域分割结果,如图 4 d 所示.

4　实验结果分析

本文选用不同背景情况的新闻视频、影视对白、演职员表等 3 类素材,每类 50 帧共 150 帧图像作为实验素材,图像尺寸为 480×360. 图 5 所示为部分测试图像的原始图像、字幕区域掩模图像及分割结果. 实验结果如表 1 所示,检到率为 96.6%,检准率为 93.1%,表明本文算法能够适应不同大小、不同位置的文字,同时对中文和英文都有很好的检测效果. 在时效性方面,由于需要计算所有子块的统计特征,计算量较大,难以进行实时处理,我们在 P Ⅳ 1.5 GHz CPU,512 MB 内存和 Windows XP 平台下,用 Matlab 6.51 开发的算法处理一幅 480×360 的图像大约需要 3 s.



a 位置不定字幕

b 单一字体字幕

c 多字体、多位置字幕

图 5　字幕检测原图、掩模及分割结果

表 1　本文算法实验结果

类型	帧数	应检子块数	实检子块数	实检正确子块数	检到率/%	检准率/%
新闻视频	50	2759	2689	2597	97.5	94.1
影视对白	50	1821	1755	1683	96.4	92.4
演职员表	50	2456	2351	2268	95.7	92.3
合计	150	7036	6795	6548	96.6	93.1

5 结 语

自动分割并识别出来视频字幕,对基于内容的视频索引检索是非常有意义的. 本文通过分析字幕区域在经过小波分解后所表现出来的明显特征,提出了基于小波变换的 K 均值聚类字幕区域分割算法:首先对图像进行两级分解并重构,然后将得到的图像分割为若干子块,提取其统计特征;最后用 K 均值聚类的方法对所有子块进行聚类分析,经过后续处理后得到字幕区域分割结果. 实验结果表明该算法正确识别率较高、效果基本令人满意. 但由于没有使用金字塔模型,本文算法并不能很好地适应过大或过小的字体,因此下一步的主要工作就是引入金字塔模型,使得算法对字体大小有更好的适应性,同时计算效率也需要进一步提高.

参 考 文 献

[1] Xie Yuxiang, *et al*. Caption detection in news video frames [J]. Computer Engineering, 2004, 30(20): 167-168 (in Chinese)
(谢毓湘, 等. 新闻视频帧中的字幕探测[J]. 计算机工程, 2004, 30(20): 167-168)

[2] Cai Bo, Zhou Dongru. Digital video caption detection and extraction techniques and implementation [J]. Journal of Computer-Aided Design & Computer Graphics, 2003, 15(7): 898-903 (in Chinese)
(蔡 波, 周洞汝. 数字视频中字幕检测及提取的研究和实现[J]. 计算机辅助设计与图形学学报, 2003, 15(7): 898-903)

[3] Li H P, *et al*. Automatic text detection and tracking in digital video [J]. IEEE Transactions on Image Processing, 2000, 9(1): 147-156

[4] Zhang Dongping, *et al*. Automatic text location in color images [J]. Journal of Zhejiang University: Engineering Science, 2005, 39(2): 229-233 (in Chinese)
(章东平, 等. 自动定位彩色图像中的文本[J]. 浙江大学学报: 工学版, 2005, 39(2): 229-233)

[5] Huang Xiaodong, *et al*. Retrieving Chinese captions in video images by wavelet transform and color clustering [J]. Computer Engineering, 2003, 29(1): 43-44 (in Chinese)
(黄晓东, 等. 用小波变换及颜色聚类提取的视频图像内中文字幕[J]. 计算机工程, 2003, 29(1): 43-44)

[6] Wu V, Nanmatha R, Risema E. Text finder: an automatic system to detect and recognize text in images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(11): 1224-1229

[7] Jain A K, *et al*. Automatic text location in images and video frames [J]. Pattern Recognition, 1998, 31(12): 2055-2076

[8] Zhuang Yueting, Liu Junwei, Wu Fei, *et al*. Automatic caption location and extraction in digital video based on support vector machine [J]. Journal of Computer-Aided Design & Computer Graphics, 2002, 14(8): 750-753 (in Chinese)
(庄越挺, 刘骏伟, 吴 飞, 等. 基于支持向量机的视频字幕自动定位与提取[J]. 计算机辅助设计与图形学学报, 2002, 14(8): 750-753)

[9] Li Zhaohui, *et al*. The application of wavelet-neural network in detection of digital video text for content-based video retrieval [J]. Journal of Guangzhou University, 2001, 15(5): 36-39 (in Chinese)
(李朝晖, 等. 小波-神经网络在视频文本自动检测中的应用[J]. 广州大学学报: 综合版, 2001, 15(5): 36-39)

[10] Dai Qingyun, *et al*. A kind of segmentation method of vehicle-license-plate images based on wavelet and mathematical morphology [J]. Journal of Image and Graphics: A, 2000, 5(5): 411-415 (in Chinese)
(戴青云, 等. 一种基于小波与形态学的车牌图像分割方法[J]. 中国图象图形学报: A 版, 2000, 5(5): 411-415)

王建宇 男,1964 年生,博士,教授,主要研究方向为计算机图形学、虚拟现实、计算机仿真.

张 峰 男,1982 年生,硕士研究生,主要研究方向为图像处理和模式识别(zhjinf82@163.com).

周献中 男,1962 年生,博士,教授,博士生导师,主要研究方向为智能信息处理、信息系统工程.

史迎春 男,1969 年生,博士后研究人员,主要研究方向为多媒体信息处理.

骆 文 男,1982 年生,硕士研究生,主要研究方向为图像处理和模式识别.