基于信息熵的自适应阈值视频镜头检测方法

陈卓夷　赵新生　赵　京
（邯郸学院计算机系　邯郸 056005）

摘　要　文中提出了一种基于信息熵的自适应阈值视频镜头检测方法。首先，利用小波变换提取图像的颜色特征和纹理特征，然后利用信息熵方法来实现对突变和渐变镜头边界的检测，并根据滑动窗口中差值的分布来动态计算局部阈值，提高了镜头边界检测算法的精度。该方法能较好地检测出镜头突变，对渐变镜头也能达到检测的目的。实验结果表明算法能够有效地检测出视频镜头边界。

关键词　视频，镜头检测，信息熵，自适应阈值

A Novel Shot Detection Method Using Automatic Threshold Based on Information Entropy

CHEN Zhuo-Yi　ZHAO Xin-Sheng　ZHAO Jing
(Department of Computer Science, Handan College, Handan 056005)

Abstract　In this paper, an efficient shot detection method approach is presented based on information entropy. Integrating color and motion information is used to descript frame content, then shot detection is accomplished based Information Entropy. The local thresholds to identify the shot transitions are computed dynamically using differrnce between consecutive frames in the sliding window. So the precision of shot detection is improved. Our method is not sensitive to brightness change and quick motion. Therefore, it can improve the precision of detecting shot boundaries. Finally we give the experimental results and draw the conclusion.

Keywords　Video, Shot detection, Information entropy, Automatic threshold

1　引言

随着多媒体技术和计算机网络技术的发展，人们接触到的视频数据以前所未有的速度增长，这就需要对这些海量的视频数据进行有效的组织和管理。由于视频数据不同于传统数据库所处理的数据类型，它不是一种简单的数值或字符型数据，因此，传统数据库中对字符或数值型数据的处理方法已经完全不能适应对视频数据的处理要求。另外，随着数字电视、视频点播(VOD)以及多媒体搜索引擎等技术的出现，仅能提供线性视频浏览方式的传统多媒体系统并不能满足这些系统在复杂网络环境下的传输、检索和浏览等功能的要求。正是在这种背景下，用来有效地分析、组织和管理海量视频数据的基于内容的视频检索技术(Content-Based Video Retrieval, CBVR)应运而生，成为目前多媒体技术的研究热点。

从内容组成的角度来说，无论什么影视节目，都是由一系列的镜头通过不同的方式利用各种剪辑手段组接而成的；反之，当给定一个视频片段，对其内容进行分析的首要步骤便是把这个视频片段沿时间轴分解成一个个以镜头为单位的单元，然后在镜头的基础上构建场景、建立视频摘要和索引结构，因此，镜头边界检测是视频分析和检索的重要基础。文中综合考虑了图像的颜色特征和纹理特征，提出了一种自适应阈值的镜头边界检测算法，根据差值的分布自动计算阈值，因此具有较高的查全率和精确度。

2　典型的镜头边界检测技术

镜头切换一般伴随着视觉内容的变化，这种变化通常表现为颜色差异增大、新旧边缘的远离、对象形状的改变和运动的不连续性等。因此，镜头边界检测的首要问题是提取适当的特征并设计帧间内容差异的度量方法，之后便可根据这个差异采用一定的策略来判断是否发生了镜头切换。在镜头边界检测方面，人们已经做过大量的研究，也提出了很多检测方法。

2.1　基于像素比较的方法

图像中各个像素的灰度或颜色是描述帧的最直接特征。像素比较是计算连续两帧对应像素间的灰度或颜色变化：

对于灰度图像：

$$D(i,i+1)=\frac{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y}|P_i(x,y)-P_{i+1}(x,y)|}{N_x\times N_y}$$

对于彩色图像：　(1)

$$D(i,i+1)=\frac{\sum_{x=1}^{N_x}\sum_{y=1}^{N_y}\sum_{c}|P_i(x,y,c)-P_{i+1}(x,y,c)|}{N_x\times N_y}$$

式中，$P_i(x,y)$和$P_i(x,y,c)$分别表示第i帧(x,y)位置像素点的灰度值和颜色值；N_x和N_y分别是图像帧的水平和垂直分辨率。如果第$i,i+1$帧的帧间差$D(i,i+1)$超过预定义的阈值，则认为此处发生了镜头切变。这种方法的缺点是对局部噪声或运动非常敏感，因为它严格地限定了像素的空间位置。摄像机的任何移动都会使帧间差异明显增大，很容易导致误检。为了减少运动造成的影响，文[1]先对相邻的图像帧使用3×3的均值滤波器，平滑处理之后再计算帧间差。但是对于较大的运动，帧间差异仍然很大而导致误判严重。

2.2　基于直方图比较的方法

直方图法是计算帧图像的直方图$H[f(x,y,t),k]$，其中$k=0,1,\cdots,N-1$，并比较连续两帧图像直方图的对应统计，目前常用的是明考斯基距离(Minkowski Distance)Lp定义为：

$$D(i,i+1)=\left(\sum_{j=1}^{N}|H_i(j)-H_{i+1}(j)|^p\right)^{\frac{1}{p}}\qquad(2)$$

其中，$H_i(j)$和$H_{i+1}(j)$分别是帧$i,i+1$的直方图在灰度(彩色)级j上的值；N是灰度(彩色)级的数量，p是正整数。此

陈卓夷　副教授，主要研究方向：多媒体信息处理；赵新生　副教授，主要研究方向：软件设计；赵　京　工程师，主要研究方向：人工智能。

外还有 x^2 直方图、直方图相交、二次型距离、EMD 距离[2]以及基于信息论的度量[3]等方法。由于直方图描述的是一幅图像像素总体的灰度或颜色分布，而忽略了像素的空间位置信息，因此，基于直方图比较的方法对小的运动和噪声不太敏感。但也正因为如此(直方图忽略了像素的空间位置信息)，有可能两幅图像内容完全不同，直方图却非常相似甚至完全相同，从而产生漏检。

2.3 基于编辑模型的方法

基于编辑模型的方法是利用镜头编辑的先验知识，对各种镜头边界类型建立一定的数学模型，自上而下地进行镜头边界的检测。

设 $f(x,y,t)$ 表示视频序列 f 在时刻 t，位置为 (x,y) 的像素值。如果视频序列在时间段 $[t_1,t_2]$ 内，完成从镜头 g 到镜头 h 的转换，则这个镜头边界可以用如下数学模型描述：

$$f(x,y,t)=(1-\alpha(x,y,t))g(x,y,t)+\alpha(x,y,t)h(x,y,t)\quad t_1\leqslant t\leqslant t_2 \tag{3}$$

其中，变换函数 $\alpha(x,y,t)$ 可以是线性的，也可以是非线性的，它定义了作为 $g(x,y,t)$ 和 $h(x,y,t)$ 混合结果的 $f(x,y,t)$ 的变化过程。如对于突变过程，有 $\alpha(x,y,t_1)=0$，$\alpha(x,y,t_2)=1$ 且 $t_2-t_1=1$；而对于叠化过程，有 $0\leqslant\alpha(x,y,t)\leqslant 1$，$\alpha(x,y,t)<\alpha(x,y,t+1)$ 且 $t_2-t_1>1$。为了减小问题的难度，现有工作都对这个模型进行了简化，文[4]提出一个统一的镜头边界检测模型，模型中分别运用 10 个参数来表征突变和渐变镜头边界。

3 自适应阈值的镜头边界检测算法

从以上典型的镜头边界检测技术的分析中可以看到，镜头分割基本上都是采用单一的判别准则来进行的。而单一的判别准则不能有效地检测镜头边界，需要将多种特征综合起来考虑。在前面介绍的技术中都需要预先确定阈值，使得算法的灵活性和适应性都受到了较大的限制，由镜头定义可知，在同一镜头内，视频帧内容相似，而在镜头边界处，视频帧内容存在较大的差异，相邻帧间将会产生整个颜色组成的显著变化或纹理信息的显著变化，或者两种情况都有。因此，我们综合考虑了图像间的颜色特征和纹理特征，提出了一种自适应阈值的镜头边界检测算法。

3.1 特征提取

小波变换不但能够实现信号的多分辨率分析，而且在时频两域中都具有表征信号局部特征的能力，即在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率，被誉为分析信号的数学显微镜。因此，小波变换具有广泛的应用领域，小波变换能够减少图像的维数，颜色特征是从小波变换的低频带提取的，纹理特征从高频带提取。

我们用离散的小波变换分解图像数据成小波系数，一个图像 $I(x,y)_{m,n}$ 的小波变换步骤如下：

令：$C_{0,m,n}=I(x,y)_{m,n}$，$m,n\in Z$

$$C_{j,m,n}=(H\oplus H)C_{j-1}=\sum_{k,l}C_{j-1,k,l}h_{k-2m}h_{l-2n} \tag{4}$$

$$L^1_{j,m,n}=(H\oplus G)C_{j-1}=\sum_{k,l}C_{j-1,k,l}h_{k-2m}g_{l-2n} \tag{5}$$

$$L^2_{j,m,n}=(G\oplus H)C_{j-1}=\sum_{k,l}C_{j-1,k,l}g_{k-2m}h_{l-2n} \tag{6}$$

$$L^3_{j,m,n}=(G\oplus G)C_{j-1}=\sum_{k,l}C_{j-1,k,l}g_{k-2m}g_{l-2n} \tag{7}$$

其中，H 和 G 为高、低通滤波器，图像 $I(x,y)_{m,n}$ 已分解成尺度为 j 的近似分量 C_j 和三个细节分量 L^1_j，L^2_j 和 L^3_j。C_j 是描述原始图像在低分辨率上的一个近似，因此从低频带提取的特征向量的近似反映颜色信息，颜色特征向量表示为 $V^C=\{C_{d,1,1},\cdots,C_{d,1,n},\cdots,C_{d,m,1},\cdots,C_{d,m,n}\}$。

纹理特征反映图象的某种局部性质，或是对局部区域中像素之间关系的一种度量。小波变换的高频统计量用于描述纹理特征，纹理特征向量能够定义为一个小波变换的三个高频矩阵总和的向量。标准偏差是：

$$D_j=\frac{1}{N}\sum_{m,n}|L^i_{j,m,n}-u^i_j|^2 \tag{8}$$

其中，i 是尺度为 j 的第 i 个细节分量，$u^i_j=\frac{1}{N}\sum_{m,n}L^i_{j,m,n}$，纹理特征向量表示为：$V^T=\{D^1_1,D^2_1,D^3_1,\cdots,D^1_d,D^2_d,D^3_d\}$

3.2 基于信息熵的帧间差计算

如果把每个图像帧都作为一个信源，那么视频序列可以看作是一系列的信源，其中每个信源都能提供一定的信息量。属于同一个镜头的各帧在内容上具有很强的相似性，考虑到视频内容的时序特点，可以认为这些信源的信源空间变化不大，因此，它们提供的信息量也不会差别很大。而当发生镜头转换时，相邻两帧的内容有较大变化，同样考虑到视频内容的时序特点，可以认为这两个信源的信源空间有较大的不同，从而使得这两帧所提供的信息量有较大的差别。

作为信源总体信息测度的确定的量，信息熵(Entropy)提供了信息量的计算方法，这里把帧图像的信息熵称为图像熵，根据上面的直观论述，可以根据帧序列图像熵的变化来判断是否发生了镜头转换。

根据 3.1 节提取的颜色特征向量 $V^C=\{C_{d,1,1},\cdots,C_{d,1,n},\cdots,C_{d,m,1},\cdots,C_{d,m,n}\}$，经归一化后的 $\overline{V^C}$ 作为局部密度分布函数 $p(s)$，计算图像熵 $H_C(s)$

$$H_C(s)=-\sum_{h=1}^{n}p_h(s)\log\ p_h(s) \tag{9}$$

帧间差即为图像 F_i 和图像 F_{i+1} 的图像熵差值：

$$Dist_C(F_i,F_{i+1})=|H_C(F_i)-H_C(F_{i+1})| \tag{10}$$

其中，$Dist_C(F_i,F_{i+1})$ 为基于颜色特征的帧间差。

同理，根据 3.1 节提取的纹理特征向量 $V^T=\{D^1_1,D^2_1,D^3_1,\cdots,D^1_d,D^2_d,D^3_d\}$，经归一化后的 $\overline{V^T}$ 作为局部密度分布函数 $p(s)$，计算信息熵 $H_T(s)$，基于纹理特征的帧间差为 $Dist_T(F_i,F_{i+1})$。

计算 $Dist_z(F_i,F_{i+1})=Dist_C(F_i,F_{i+1})+Dist_T(F_i,F_{i+1})$，$Dist_z(F_i,F_{i+1})$ 为融合了颜色特征和纹理特征的帧间差。

3.3 检测镜头边界

如果镜头发生变化，两帧间的帧间差 $Dist_z(F_i,F_{i+1})$ 将超过相应的阈值 T_Z。下面给出一种简单的自适应阈值方法来检测镜头边界，设置一个大小为 W 的滑动窗口来处理连续 W 帧间的比较值，在滑动窗口 W 中，我们为帧间差 $Dist_z(F_i,F_{i+1})$ 计算其均值 $\overline{Dist_z(F_i,F_{i+1})}$，利用这个均值计算自适应阈值 T_Z(均值的 4 倍)。完成了上述检测后，清空滑动窗口，继续下一个 W 帧的检测，如果剩余的帧数小于 $W/2$，则计算剩余所有的帧，滑动窗口 W 的大小是个经验数值，一般设置为 300 视频帧，这样可以在一定程度上消除噪声的影响，增加阈值的鲁棒性。

在滑动窗口内有孤立的峰表明该处有突变，如果没有孤立峰，但存在连续帧的帧间较大差值，则判断该处为渐变镜头边界，如图 1 所示。

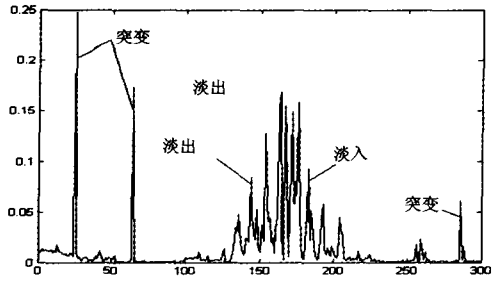

图 1 本算法的检测结果

4 实验

镜头边界检测结果的评价包括查全率(Recall)和查准率(Precision)两个指标,它们的定义分别为:

查全率=正确检测数/(正确检测数+漏判数)

查准率=正确检测数/(正确检测数+错判数) (11)

这两个指标往往是一对矛盾:仅仅为了提高查全率,通常会导致误检,使得查准率下降;而仅仅为了提高查准率,通常会导致漏检,使得查全率下降。因此,查全率和查准率都高才能说明检测方法好。

在实验中,选用了五个具有不同特点的视频片段组成本系统的实验数据集,包括:电影《简爱》中聚会的片段(JE)、儿童片《绿野仙踪》中庆祝坏女巫死亡的片段(LY)、体育比赛中"排球比赛"的结尾片段(PQ)、晚间新闻中的反映贫困山区教育片段(XW)和 MTV《赞酒歌》中歌伴舞片段(ZJ)。视频序列从 4 千多帧到 2 万多帧,每帧为 352×240 像素。

表 1 实验结果

视频片段	帧数	突变	渐变	本算法(%)		VideoAnnEx 算法(%)	
				查准率	查全率	查准率	查全率
JE	18644	76	6	89.9	92.2	95.1	80.3
LY	11740	38	3	90.5	94.2	94.6	82.7
PQ	4576	23	0	92.6	95.5	98.6	74.7
XW	4650	16	0	88.4	93.3	91.9	98.3
ZJ	25532	44	3	85.7	95.8	93.9	83.3

表 1 给出了本算法与 VideoAnnEx 系统中镜头边界检测算法[5]在突变和渐变镜头边界检测中的实验对比结果。从实验结果可以看出,文[5]算法虽然对运动和闪光灯场景具有较高的鲁棒性,但同时取得了较低的查全率。从视频结构分析的角度来看,由于镜头边界检测是场景分析的基础,在场景分析中可以有效地合并过分割的镜头,因此,查全率相对于查准率更为重要。本文方法既取得了较好的查准率,同时也取得了较高的查全率。

图 2 中的实验结果取自于电影《简爱》中 300 视频帧,图 2(a)给出的是基于颜色特征的图像熵差值的变化的曲线,检测出包含 5 个切变,其中错检 2 个,漏检 1 个。图 2(b)给出的是利用本算法融合了颜色特征和纹理特征的图像熵差值变化的曲线,检测出 4 个切变,与人工检测的结果完全一致,因此避免了镜头内物体运动或光线突然发生变化时而发生的误检。

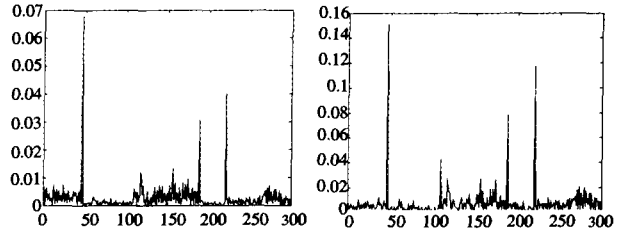

图 2 图像熵差值变化曲线

结论 文中介绍了典型的镜头边界检测的基本方法,并针对其不足提出了一种基于信息熵的自适应阈值的镜头边界检测算法,首先利用小波变换提取了帧图像的颜色特征和纹理特征,然后利用信息熵的方法来计算帧间差,由于融合了颜色特征和纹理特征,从而避免了镜头内物体运动或光线突然发生变化时而发生的误检,并根据滑动窗口中差值的分布来动态计算局部阈值,该方法能较好地检测出镜头突变,对渐变镜头也能达到检测的目标,下一步可以对分割出来的镜头进行语义上的分析聚类,形成高层的视频结构,以提供视频基于语义的检索。

参 考 文 献

1 Zhang H J, Kankanhalli A, Smoliar S W. Automatic partitioning of full-motion video. ACM Multimedia Systems, 1993, 1(1): 10～28

2 Rubner Y. Perceptual metrics for image database navigation: [PhD Thesis]. Stanford University, May 1999

3 Cheng W G, Xu D, Jiang Y W, Lang C Y. Information theoretic metrics in shot boundary detection. In: Proc. of IIH-MSP, LNCS 3683, Melbourne, Australia, Sept. 2005. 388～394

4 Bescos J, Cisneros G, et al. A unified model for techniques on video-shot transition detection. IEEE Trans. on Multimedia, 2005, 7(2): 293～307

5 Amir A, Berg M, et al. IBM research TRECVID-2003 video retrieval system. In: TRECVID Workshop, Washington D. C. USA, Nov. 2003

(上接第 225 页)

插值平滑技术。在实验中我们发现,Chen-Goodman 方法比其他方法优越的原因在于该方法所引发的词性聚类倾向比较明显。词性因素可能是划分模型参数的重要因子。基于这种启发,我们对各种方法导致的词性聚类作了对比分析,并在此基础上,给出了 2 种改进的平滑方法。实验结果表明,这 2 种方法比原来的方法有相对更好的适应性。

参 考 文 献

1 Chen S F, Goodman J. An Empirical Study of Smoothing Techniques for Language Modeling: [Technical Report TR-10-98]. Computer Science Group, Harvard University, 1998

2 Gale W A, Church K W. What's wrong with adding one? In: N. Oostdijk, P. de Haan, eds. Corpus-Based Research into Language. Rodolpi, Amsterdam, 1994

3 Gale W A, Sampson G. Good-Turing frequency estimation without tears. In: Journal of Quantitative Linguistics, 1995, 2(3): 217～237

4 Church K W, Gale W A. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. Computer Speech and Language, 1991, 5(1): 19～54

5 Chen S F. Building probabilistic models for natural language. Harvard University, Cambridge, MA, 1996

6 Jelinek F, Mercer E L. Interpolated Estimation of Markov Source Parameters from Sparse Data. In: D. Gelsema and L. Kanal, eds. Pattern Recognition in Practice. North-Holland, 1980

7 Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE ASSP, 1997, 35 (3): 400～401

8 Kneser R, Ney H. Improved Backing-off for M-Gram Language Modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1995, 1: 181～184

9 Ney H, Essen U. On smoothing Techniques for Bigram-based Natural Language Modelling. In: Proceedings of the IEEE 1991 International Conference on Acoustic, Speech, and Signal Processing, Toranto, 1991. 251～258

10 Manning C D, Schutze H. 统计自然语言处理基础. 苑春法, 等译. 电子工业出版社, 2005