结合互信息量与模糊聚类的关键帧提取方法①

蔡家楣 陈 洋 陈铁明 张旭东 (浙江工业大学 软件学院 浙江 杭州 310023)

摘 要： 关键帧是描述一个镜头的关键图像帧，它通常反映一个镜头的主要内容，因此，关键帧提取技术是视频分析和基于内容的视频检索的基础。提出了一种结合互信息量与模糊聚类的关键帧提取方法，一方面通过互信息量算法对视频片段进行镜头检测可以保持视频的时间序列和动态信息，另一方面通过模糊聚类使镜头中的关键帧能很好的反映视频镜头的主要内容。最后构建了一套针对 MPEG-4 视频的关键帧提取系统，通过实验证明该系统提取的关键帧，可以较好地代表视频内容，并且有利于实现视频分析和检索。

关键词： 互信息量；镜头检测；模糊聚类；关键帧提取；视频检索

Combination of Mutual Information and Fuzzy Clustering for Key-Frame Extraction

CAI Jia-Mei, CHEN Yang, CHEN Tie-Ming, ZHANG Xu-Dong

(College of Software Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: A key-frame is a representation of one shot's content in the video, which usually reflects the main elements of a scene. Therefore, key-frame extraction and video analysis technology is the basis of content-based video retrieval. In this paper, a key-frame extraction method based on combination of mutual information and fuzzy clustering is proposed. In this method, the key-frames can maintain time-series and dynamic information of the video. And also, the key-frames can be a good reflection of the main contents of the video. Finally, one key-frame extraction system for MPEG-4 video is designed, and experiments show that the key-frame extraction system can be a better representative of video content, and conductive to the realization of video analysis and retrieval.

Keywords: mutual information; shot detection; fuzzy clustering; key-frame extraction; video retrieval

1 引言

关键帧提取技术是视频语义检索的关键技术，是视频信息索引、浏览与检索的基础。经过各研究机构的多年研究，目前已经取得了不少研究成果，有基于镜头边界法、基于内容分析法、基于运动分析法和基于视频聚类法等。在这些方法中，基于模糊聚类算法提取的关键帧能很好的反映视频的主要内容，并有效的消除镜头间的相关性，但其缺点是不能保持视频信息的时间序列和动态信息，不利于视频索引建立与内容分析。本文提出了一种结合互信息量和模糊聚类的关键帧提取方法，首先利用连续帧间的互信息量检测视频序列的镜头边界，将视频片段划分为若干个子镜头，然后在镜头内采用模糊聚类算法提取出能有效反映本镜头内容的关键图像帧。在此基础上，针对大多数视频以 MPEG-4 压缩格式存在的现实情况，本文构建了一套简便有效的关键帧提取系统，通过实验表明，该系统提取的关键帧具有代表性和有效性，同时有效的保持了原始视频图像帧的时间顺序和动态信息，有利于实现视频分析和检索。

2 系统总体结构

图 1 是本文构建的基于内容的关键帧提取系统。

① 基金项目:浙江省信息产业厅项目(K0853119001900)

收稿时间:2009-07-12;收到修改稿时间:2009-09-19

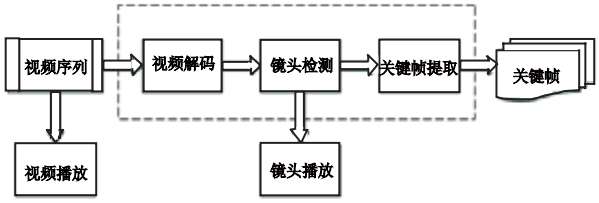

图 1　系统总体结构图

从图 1 中可知，系统主要由五个部分组成：视频解码、镜头检测、关键帧提取、视频播放和镜头播放，其中前三个模块为视频核心模块。视频播放和镜头播放是系统的辅助功能，前者负责对整个视频序列的播放，后者主要针对某个镜头片段的播放。核心模块中，视频解码首先对新加入的 MPEG-4 压缩视频流进行解码，提取每帧图像的 YUV 颜色特征值[1]，其次对整个视频序列进行分段得到视频镜头集，然后对每个镜头进行关键帧提取，得到的视频帧代表这个镜头的主要内容。

2.1 视频解码

MPEG-4 对视音频编码的结果数据率支持在一个带宽范围内可适应性变化，使总带宽为 5Kbps~4Mbps；它能适应视频、图片基于内容的可伸缩性，在空域和时域对图像质量的可伸缩性，可实现对视频序列和各种图片可扩展操作处理，也可实现针对内容对象的随机访问[2]。基于这些好的特性，大量的视频数据都是以 MPEG-4 等压缩形式存在，因而有必要在视频分析之前对压缩视频进行解码处理。

在视频解码过程中，将整个图像画面看作一个对象，即视频对象平面 VOP(Video Object Plane)，调整分辨率为 CIF(Common Intermediate Format, 352×288 像素)，提取 YUV 颜色特征值，按照帧的索引值，以文件的方式分别存储 Y、U 和 V 三个分量值。选用 YUV 颜色特征原因主要有两方面：1)MPEG-4 大多支持 YUV420 平面模式;2)YUV 能较好的代表原视频图像的内容，并且可以方便的转换为其它格式像素值。

2.2 基于互信息量和联合熵的镜头检测

镜头检测是视频处理的前提，是基于内容的视频检索的关键技术。镜头是指摄像机在一个连续的时间和空间中拍摄得到的视频序列，它由一系列连续帧组成的，一般而言，同一镜头内各帧之间差异值较小，而不同镜头的帧之间差异较大。根据互信息量的相关理论[3]，相邻帧间图像 t,t+1 的互信息量 MIH(t,t+1)可以定义为两帧图像在同一灰度级上像素分布相关性的量度。在本文所采用的镜头检测算法中，将帧间的互信息量 MIH(t,t+1)作为特征参数参与帧间变化的计算，能有效地描述帧间变化。设一段包含 N 帧的视频序列为 $F=\{f_1,f_2,\cdots,f_n\}$，镜头边界检测算法主要步骤如下：

1) 计算相邻帧间相似度

互信息量的定义表明：两帧图像内容改变较大时，相应的互信息量较小，而当两者内容基本相同时，互信息量则会保持一个较高值。也就是说，互信息量的高低可以反映帧间相似度。定义帧间相似度 Dt 的计算公式如下：

$$D_t=MIH_{t,t+1}*(1-Diff_{t,t+1}) \quad (1)$$

$MIH_{t,t+1}$ 代表两帧图像之间的互信息量，为了减少计算量，将图像的灰度级别调整为 32 色。$Diff_{t,t+1}$ 代表两帧图像 t,t+1 之间的 HSV 颜色直方图特征差异值(颜色直方图是常用的图像颜色特征表示方法，反映图像颜色的统计分布，描述的是图像的整体颜色特征)，将其作为帧间相似度的权值，可以消除运动带来的误检。

2) 帧间相似度差值

针对所得到的帧间相似度，若采用不变的阈值来判断某处是否存在可能的镜头切换，势必会降低算法的性能。利用帧间相似度差值可以较为容易的确定镜头切换位置，在计算中每 500 帧选取一次阈值 T_s[4]，则帧间相似度之差大于 T_s 的镜头都是可能的镜头切换，放入可能镜头集合 S_{ps}。帧间相似度差值 $D1_t$ 计算公式为：

$$D1_t=(|D_t-D_{t-1}|+|D_{t+1}-D_t|)/2 \quad (2)$$

3) 去除噪声

由于可能受到闪光灯、运动等因素干扰，镜头切换点并不能代表真实的镜头变化，也就需要去除这些潜在的噪声。遍历检测相邻帧的变化值，如果出现连续两次帧间差异值都小于一个较小阈值 T_e，那么他们之间的差异很可能只是受到闪光灯影响，起始帧以及之后的连续 3 帧都被排除出镜头集合 S_{ps}。同样，如果镜头起始帧和下一帧的信息量差值不为最小，那么可能只是噪声干扰，其起始帧并不为镜头切变点。

4) 动态窗口

镜头的切变过程可以在连续两帧之间完成，而渐变过程则可能持续几十帧，为了将两者统一进行检测，采取动态窗口来描述完整的镜头切换过程。对镜头集合 S_{ps} 中的每一个位置 t，首先计算出 t 为起始帧的固定窗口 W_C，窗口 W_C 的结束点为 $t+W_C$ 帧。然后在 W_C 内选取一动态窗口 W_D，并计算动态窗口 W_D 内每一帧与 t-1 帧间的相似度，找到数据平稳的开始处作

为 W_D 的结束点。

5）区分切变镜头和渐变镜头

根据帧间互信息量的特性，它的值越小，说明两帧图像的内容越不相似，也就意味着它是真正的镜头切换的可能就越大。所以在可能镜头集合 S_{ps} 中，只需要通过计算其动态窗口的大小就可以区分切换点是切变过程还是渐变过程，如果动态窗口的大小为 1，那么该切换点就是切变，否则如果大于 1，那么该切换点就是渐变过程。

2.3 基于模糊聚类的关键帧提取

FCM 算法[5,6]是一种基于划分的聚类算法，目前已经在诸多领域获得了广泛的应用，并取得了满意的效果。它的主要思想就是使得被划分到同一簇的对象之间相似度最大，而不同簇之间的相似度最小。对于镜头集合 $S=\{S_1,S_2,\cdots,S_n\}$，本文采用模糊聚类算法提取镜头 S_i 的关键帧，基本原理为：在镜头内部根据图像熵进行聚类，提取最接近聚类中心的帧作为关键帧，而帧的数量主要取决于镜头内类的数量。

为了提高聚类的效率，本文选择 YUV 颜色空间中灰度级为 32 色的颜色直方图作为视觉特征，以欧几里德距离计算两个视频帧之间的相似度。具体的步骤如下：

1）初始化参数

设置样本点个数(镜头的帧数)，每个样本点的维数(32*3)，要聚类的数目(每 80 帧提取一帧)，参数 m (根据经验值设为 2)以及误差限(0.1)等，读取镜头内每帧图像 YUV 颜色直方图。

2）初始化聚类中心

FCM 用模糊划分隶属度的归一化规定，一个数据集的隶属度 u 的和总等于 1：

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1,...,n \tag{3}$$

用值在 0，1 间的随机数初始化隶属矩阵 U，使其满足公式(3)中的约束条件，并将随机选取的初始迭代点作为初始的聚类中心。

3）计算划分矩阵

根据欧几里德距离计算每帧图像与聚类中心的距离，将镜头内的帧划分到各个矩阵。

$$J(U,c_1,...,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{4}$$

这里 u_{ij} 介于 0，1 间；c_i 为模糊组 i 的聚类中心，$d_{ij}=||c_i-x_j||$为第 i 个聚类中心与第 j 个数据点间的欧几里德距离；且 $m\in[1,\infty)$是一个加权指数。

4）构造新的目标价值函数

为了求得每组的聚类中心，需要使得非相似性(或距离)指标的价值函数达到最小。故构造如下新的目标函数，可求得使(4)式达到最小值的必要条件：

$$\begin{aligned}\overline{J}(U,c_1,...,c_c,\lambda_1,...,\lambda_n) &= J(U,c_1,...,c_c) + \sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1) \\ &= \sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^m d_{ij}^2 + \sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1)\end{aligned} \tag{5}$$

这里 λ_j(j=1…n)，是(3)式的 n 个约束式的拉格朗日乘子。对所有输入参量求导，使公式(4)达到最小的必要条件为：

$$c_i = \sum_{j=1}^{n} u_{ij}^m x_j \Big/ \sum_{j=1}^{n} u_{ij}^m \tag{6}$$

和

$$u_{ij} = 1 \Big/ \sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)} \tag{7}$$

记录目标价值函数的结果值，作为循环迭代执行的判断条件。

5）重新计算聚类中心

根据新的隶属 U 矩阵，重新计算聚类中心矩阵。如果目标价值函数的值小于某个确定的阈值，或相对于上次价值函数值的改变量小于某个阈值，则算法停止。为了避免不良初始聚类中心的危害，本文加入了一个极限迭代值，如果达到这个值，算法也将停止。否则返回步骤 3)。

根据以上算法步骤，最后将输出 C 个聚类中心点向量和 C*N 的一个模糊划分矩阵，这个矩阵表示的是每个图像帧属于每个类的隶属度。根据这个划分矩阵按照模糊集合中的最大隶属原则就能够确定每个图像帧归为哪个类，镜头内的离聚类中心最近的图像帧作为关键帧。

3 实验与分析

根据以上的系统设计方案，本文采用 VC++6.0 开发了视频关键帧提取系统，在实验测试中，考虑到取材的广泛性和普遍性，分别选取了运动、动画、新闻和广告四种类型的视频片段，检查系统的效果。

本文采用查全率和查准率这两个参量来对系统中基于互信息量的镜头检测算法，镜头检测的查全率和查准率的定义：查全率=正确检出数/(正确检测数+漏检数)；查准率=正确检测数/(正确检测数+误检数)。

根据本文提出镜头采用模糊聚类提取关键帧的方法，漏检将会影响镜头内所提取关键帧的质量，误检

会增加所提取关键帧的数量，因此查全率和查准率越高说明检测的效果越好。对比基于颜色直方图的镜头检测算法[7]的实验结果数据如表 1 所示：

表 1　镜头检测结果

视频类型	检测算法	检出数	误检数	漏检数	时间(s)
运动片段	互信息量	31	3	2	38
(1468)	颜色直方图	28	2	5	6
动画片段	互信息量	21	3	1	30
(1030)	颜色直方图	18	4	4	4
新闻片段	互信息量	32	1	1	33
(1135)	颜色直方图	28	3	5	5
广告片段	互信息量	7	1	0	5
(204)	颜色直方图	5	0	2	1

从以上实验数据可知，基于互信量和联合熵的镜头检测算法在查全率(95.8%)和查准率(91.9%)上都有较好的表现，能够很好的满足视频镜头检测的要求，但是由于算法的复杂度，在时间效率方面稍逊一筹。

为了检测系统中关键帧提取的效果，采用了国外视频检索研究(http://www.vide okeyframes.de)的视频片段 blade_run ner.mpg 作为实验测试视频，视频总共为 571 帧，分辨率为 352*144。图 2 是文献[8]基于视频特征和多边形简化提取的关键帧。

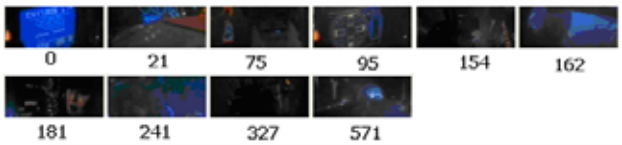

图 2　文献[8]提取的关键帧

利用本文所提出的方法提取出关键帧如图 3 所示。

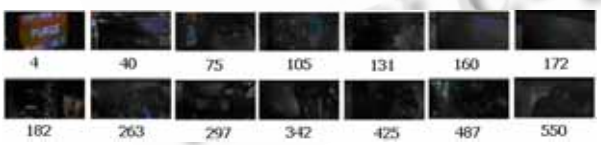

图 3　本文提取的关键帧

从图 2 和图 3 看，文献[8]把视频的首帧和尾帧都作为了关键帧，并且关键帧的分布不是很均匀，本文提取的关键帧在内容上更能够代表视频的内容，并且关键帧的分布均匀，保留了视频序列的动态信息。

因此，从总体上看，本文构建的关键帧提取系统能够较好的完成视频解码、镜头检测和关键帧提取的任务。结合互信息量与模糊聚类的关键帧提取方法，一方面发挥了互信量理论在镜头检测中的良好特性，效果显示了较好的查全率和查准率，并且有效的保持了视频序列的时间顺序和动态信息，另一方面模糊聚类算法提取的关键帧能很好的反映视频内容。

4　结语

本文提出了一种结合互信量与模糊聚类的关键帧提取方法，并以此设计和实现了一个关键帧提取系统。实验结果表明能够较好的完成的关键帧提取任务，且检测效果较好，有利于视频索引建立和内容分析。但由于算法的复杂度增加，所以系统的提取效率较低。如何提高系统的提取效率，是下一步工作的重点。

参考文献

1 黄学军,邢爱凤,解培中.综合利用边缘和颜色特征的图像检索.南京邮电学院学报(自然科学版), 2004,24(1):27－30.

2 贺贵明,吴元保,蔡朝晖,蒋旻,刘振盛,刘志雄.基于内容的视频编码与传输控制技术.武汉:武汉大学出版社, 2005.

3 Mentzelopoulos M, Psarrou A. Key-frame extraction algorithm using entropy difference. International Multimedia Conference: Proc. of the 6th ACM SIGMM international workshop on Multimedia information retrieval, 2004. 39－45.

4 Zhu XQ, Xue XY, Wu LD. An autom atic threshold detection method in video shot segmentati on. Jounal of Computer Research and Development, 2000:80－85.

5 Lo CC, Wang SJ. Video segmentation using a histogram-based fuzzy c-means clustering algorithm. Computer Standards&Interfaces. 2001. 429－438.

6 朱映映,周洞汝.一种基于视频聚类的关键帧提取方法.计算机工程, 2004,30(4):12－13.

7 Browne P, Smeaton AF, Murphy N, et al. Evaluation and combining digital video shot boundary detection algorithms. Proc. of the Fourth Irish Machine Vision and Information Processing Conference, Queens University Belfast, 2000. 136－148.

8 Latecki LJ, Wildt DD, Hu JY. Extraction of key frames from videos by optimal color composition matching and polygon simplification. Multimedia Signal Processing. 2001. 245－250.