基于非参数密度估计聚类的关键帧提取方法

陈卓夷
(邯郸学院计算机系　河北邯郸 056005)

摘　要　关键帧提取是基于内容的视频检索的一个重要的组成部分，所提取的关键帧的有效性，直接影响视频检索的结果。文中提出了一种基于非参数密度估计聚类的关键帧提取方法。首先，通过提取图像的颜色特征和运动特征，然后利用均值漂移聚类方法对融合了颜色和运动信息的特征空间进行聚类。它能自动确定类别数并具有严格的收敛性，从而大大减少了运算量，提高了运算速度。实验证明，本方法的提取结果与人的主观视觉感知系统具有良好的一致性。

关键词　视频摘要，关键帧提取，均值漂移，非参数聚类

Key-Frame Extraction Using Nonparametric Clustering Based on Density Estimation

CHEN Zhuo-Yi
(Department of Computer Science, Handan College, Handan Hebei 056005)

Abstract　Key-frame extraction has been recognized the important research issue in content-based video analysis. In this paper, an efficient key-frame extraction approach is presented based on nonparametric clustering, which provides the capability of browsing digital video sequences more efficiently. Integrating color and motion information is used to descript frame content, then key-frame extraction is accomplished by density-estimation-based nonparametric clustering, the mean shift method, which can efficiently analyze complex multimodal feature space and delineate arbitrarily shaped clusters in it. An experimental system has been build up. Experiments verify the effectiveness of the proposed approach.

Keywords　Video abstraction, Key-frames extraction, Mean shift clustering, Nonparametric clustering

1　引言

随着多媒体技术、数字电视和网络的发展，产生了大量的视频文档，如何对其进行管理和利用就成为迫切需要解决的问题，而基于内容的视频存取则成了对其进行有效处理的一个基础。对于一个视频文档，如电影故事片，用户可能需要在短时间内，无须浏览整部影片而仅了解其主要内容即能决定是否值得进行详细的观赏，因此如何快速地浏览视频，并且了解其主要内容已经成为一个重要的研究课题。视频摘要(Video Abstraction)，是解决这种问题的一个途径。除此之外，视频摘要还提供了非线性浏览视频的方式，同时它也能辅助建立视频检索和索引系统，因此对于视频摘要的研究有着重要的意义。

所谓视频摘要，是对长视频文档的简短内容总结，在表现形式上，它是一个静止或者运动的图片序列；在本质上，它是原始视频数据的一个子集。有两种基本形式的视频摘要：视频总结(Video Summary)和视频预览(Video Skimming)。由于视频总结需要通过提取一些最能反映视频内容的帧(关键帧)来组成，因此提取和生成这些帧成为视频总结的核心。本文旨在通过提取有效的关键帧集合来建立视频总结。

正是由于关键帧的提取在基于内容的检索，以及基于内容建立视频摘要中具有如此重要的地位，近年来受到了研究者的广泛关注，与之相对应地产生了一些关键帧提取算法。在早期的工作中，每隔一定时间抽取一帧作为关键帧。然而，视频内容的变化并不是匀速的，这种方法容易导致一些短小但比较重要的片段可能没有相应的关键帧，相反一些比较长的片段则可能有很多相似内容的帧。文[1]是比较当前帧与所有已提取关键帧之间的差异程度，从而确定当前帧可否作为关键帧，这种方法的结果使得各个关键帧的内容之间最大限度的不同，从而更有代表性。此方法可以根据镜头内容的变化程度选择相应数目的关键帧，但是选取的帧不一定具有代表意义，而且在有物体快速运动时，容易选取过多的关键帧。文[2]采用基于运动分析的方法，引入了全局运动的估计，从而使得选择的关键帧能够比较准确，但计算量也往往是令人生畏的。文[3]提出了一种层次凝聚类算法，首先将 N 个个体各自看成一类，然后设定样本之间的距离和类与类之间的距离。开始时各个样本自成一类，则类和类之间的距离与个体之间的距离是相等的，选择距离最近的一对合并为一个新类，计算新类和其他类的距离，再将距离最近的两类合并，这样每次减少一类，直至所有的个体都归为一类或达到结束的阈值时为止。但是它的计算量比较大，结束的阈值也很难确定。最近，Kin-wai S 等人提出[4]一种采用全局最优估计的方法提取关键帧，通过计算每个像素在镜头内相应位置点的出现概率来描述关键帧内容。此外，文[5]提出基于对象的关键帧提取方法，虽然对象可以很好地描述图像的语义内容，但目前仍然没有有效的语义对象自动提取方法。

虽然在关键帧提取上已有了许多研究成果，但寻求能涵盖整个镜头内容的关键帧仍是一个难题。所以本文提出了一种基于非参数密度估计聚类的关键帧提取算法。

2　关键帧提取算法

2.1　特征提取

镜头是视频检索的最小单位，视频分割成镜头后，就要对各个镜头进行特征提取，得到一个尽可能充分反映镜头内容

陈卓夷　副教授，主要研究方向为多媒体信息处理。

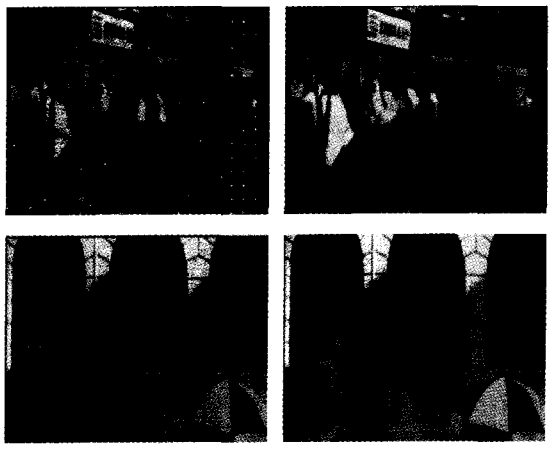
的特征空间，这个特征空间将作为视频聚类和检索的依据。为全面描述图像内容，有效提高检索性能，我们将颜色特征和运动特征结合使用构成综合特征空间。

颜色特征有较强的辨识能力，所以广泛地用来表示图像内容，本文使用颜色直方图来描述帧的视觉内容。由于 HSV 颜色空间与人的视觉感知系统有较好的一致性，本文选用 HSV 颜色空间来表示帧的颜色分量，按照人的视觉分辨能力，把 H 分成 8 份，S 和 V 各 3 份。

运动特征是分析视频数据的一个重要的且有效的线索。视频的内容通常涵盖了从高到低各种不同的活跃程度，因此，需要一个标准来准确地度量其中某个给定视频片段的运动活跃性。为此 MPEG-7 定义了一个描述符——运动活力(Motion Activity)，它直观地描述了视频片段的“动作强度”或“动作步调”。运动活力描述符提供了运动的强度、方向、空间分布、空间位置、时间分布等内容。其中，活力强度(Intensity of Activity)是一帧中运动矢量幅度的标准差，它表明了运动矢量幅度的一致性，其突出的性能已经通过实验验证，我们使用活力强度表征运动活力。运动矢量可以从 MPEG 压缩视频中直接得到，本文实验中，运动矢量通过帧间块匹配计算得到：

$$MAD(x,y)=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}|F_n(i,j)-F_{n-1}(i+x,j+y)| \tag{1}$$

其中，F_n 和 F_{n-1} 分别为当前块和前一帧中的对应块，大小均为 $N\times N$。图 1 是两个帧的镜头活力强度计算。



(a) (b)

图 1 根据运动矢量估计运动活力

2.2 关键帧提取

目前，从视频序列中检测镜头的技术已经趋于成熟。同一镜头中，有很多相似的帧，为了更加简单有效地表示镜头内容，一般提取一个或多个关键帧来表示镜头的内容。给定一个镜头内帧的集合 $Frames=\{Frame_1, Frame_2, \cdots, Frame_M\}$，关键帧的提取即确定 $Frames$ 的一个子集 $KF=\{kf_1,kf_2,\cdots,kf_N\}$，使得 KF 能表示 $Frames$ 的主要内容(其中 $M\geqslant N$，M 是一个镜头所包含的帧数，N 是提取结束后得到的一个镜头中关键帧总数)。

如今已有很多自动提取关键帧的方法，从概括镜头内容的角度来说，基于聚类的算法效果比较好，在各种聚类方法中，K 均值算法具有计算复杂性较低的优点，但是由于它需要预先设定一个类别值，并且它假定了特征空间的高斯分布(或混合高斯分布)性质，限制了它的应用。作为一种非参数的方法，层次聚类方法虽然避免了 K 均值聚类方法的弱点，但是它的计算量比较大，结束的阈值也很难确定。而均值漂移聚类(mean shift clustering)[6]算法是一种非常直观的统计迭代算法，它使每个待处理的点“漂移”到分布密度函数 $f(x)$ 的局部极大值点处，因此，它能够自动确定类别数，并且其非参数的本质导致了它在估计局部密度最大点的过程中没有强加给特征空间特定的结构，特征空间的分析更为准确。由于其特征空间分析的可靠性和健壮性，这种方法已经在人脸跟踪和图像分割中得到了应用。它还具有严格的收敛性，使得聚类效果更为稳定。具体的方法介绍可以参见文[6]。我们将它引用来处理视频数据。

均值漂移聚类(mean shift clustering)是一种非参数密度估计方法，它对先验知识要求最少，它完全依靠训练数据进行估计，而且可以用于任意形状密度函数的估计，因此它得到了广泛的应用，其中，基于核函数的密度估计是最常用的非参数估计方法之一。该方法是从核函数 $K(x)$(又称为窗口函数)出发，对每个样本 X_i，用正定矩阵来刻画总体 X 在 X_i 周围的局部空间结构。给定样本点 $X_i\in Rd, i=1,\cdots,n$，则随机变量 x 的密度函数的估计是：

$$\hat{f}(x)=\frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{2}$$

$$K(X)=\frac{C}{h_c^2h_m^2}k(\|\frac{X^c}{h_c}\|)k(\|\frac{X^m}{h_m}\|) \tag{3}$$

其中，h 是带宽；$K(x)$ 是核函数，一般有两种形式的核函数：径向对称核和多元正态核。当使用正态核时，第 j 个均值漂移矢量可以表示为：

$$m_{h,N}=y_{j+1}-y_j=\frac{\sum_{i=1}^{n}x_i\exp(\|\frac{x-x_i}{h}\|^2)}{\sum_{i=1}^{n}\exp(\|\frac{x-x_i}{h}\|^2)}-y_j \tag{4}$$

关键帧提取算法分以下步骤进行：

(1)对镜头内各帧 $\{F_i\}_{i=1,2,\cdots,k}$，提取特征集合 $X=\{X^c, X^m\}$ 中的特征矢量，其中，X^c 是颜色特征向量，X^m 是活力强度向量。

(2)对每个特征点 X_i 执行均值漂移过程，直至收敛得到收敛点 Z_i，存储所有的收敛点 $\{Z_i\}_{i=1,2,\cdots,M}$。

(3)对收敛点进行合并，将距离不超过带宽 $h=\{h_c,h_m\}$(h_c 是对应特征子空间 c 的带宽)的所有收敛点 Z_i 组合在一起，形成收敛点的类别 $\{C_p\}_{p=1,2,\cdots,m}$。

(4)设定一个类别中的最小帧数 MinFC，将少于 MinFC 的类别合并到邻近类别之中(或者认为是噪声，直接去掉)，最终形成 $\{C_j\}_{j=1,2,\cdots,n}$。

(5)从每个帧的类别中选取一帧 kf_j，从而得到关键帧集合 $KF=\{kf_1,kf_2,\cdots,kf_N\}$，其中 kf_j 代表了一个内容相似的帧的集合。

$$kf_j=\{Frame_k|Z_k(x_k)\in C_j\}\quad(1\leqslant k\leqslant M, 1\leqslant j\leqslant N) \tag{5}$$

聚类过程中需要经验性地设定带宽，但与层次聚类算法相比较，带宽值的确定要容易得多。在本文的实验中，使用上述时空特征取得了良好的效果，提取的关键帧能够涵盖镜头所要表达的主要内容。

3 实验结果

3.1 实验数据集

为了验证上述算法的有效性，我们实现了一个关键帧提

(下转第 162 页)

2 Cabalar P, Pearce D, Valverde A. Reducing Propositional Theories in Equilibrium Logic to Logic Programs. In:12th Portuguese Conference on Artificial Intelligence, Covilha, 2005
3 Chagrov A, Zakharyaschev M. Modal Logic. Oxford: Clarendon Press, 1997
4 Ferraris P, Lifschitz V. Mathematical Foundations of Answer Set Programming. In: We Will Show Them! Essays in Honour of Dov Gabbay, Vol 1. Artemov S, Barringer H, Garcez A, Lamb L, and Woods L, College Publications, 2005. 615～664
5 Gödel K. Zum intuitionistischen Aussagenkalkül. Anz Akad Wiss Wien, 1932, 69: 65～66
6 Hähnle R. Complexity of Many-Valued Logics. In:31st International Symposium on Multiple-Valued Logics, Warsaw, 2001
7 Lifschitz V, Pearce D, Valverde A. Strongly Equivalent Logic Programs. ACM Transactions on Computational Logic, 2001, 2(4): 526～541
8 Pearce D. A New Logical Characterisation of Stable Models and Answer Sets. Non-Monotonic Extensions of Logic Programming, Bad Honnef, 1996
9 Takeuti G, Titani T. Intuitionistic Fuzzy Logic and Intuitionistic Fuzzy Set Theory. Journal of Symbolic Logic, 1984, 49(3): 851～866

(上接第 120 页)

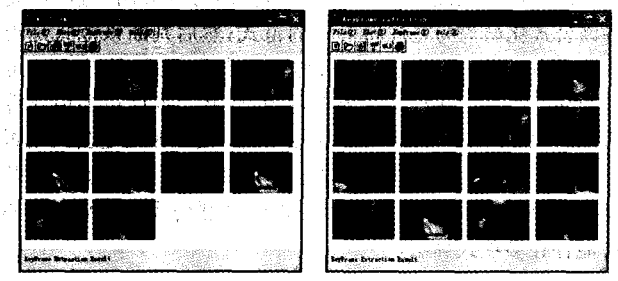
取系统,选用了五个具有不同特点的视频片段组成本系统的实验数据集,包括:爱情片《邂逅》中相聚的片段(XH)、儿童片《绿野仙踪》中庆祝坏女巫死亡的片段(LY)、体育比赛中"篮球比赛"的结尾片段(LQ)、晚间新闻中的反映植树造林片段(XW)和 MTV《赞酒歌》中歌伴舞片段(ZJ)。视频序列从 4 千多帧到 2 万多帧,每帧为 352×240 像素,其基本情况如表 1 所示。

表 1 实验数据集

视频片段	帧数	镜头数	本文算法检测出的关键帧数	算法 B 检测出的关键帧数	人工检测的关键帧数
XH	25430	66	85	86	85
LY	17928	73	99	101	99
LQ	9010	29	47	48	47
XW	4558	20	41	41	41
ZJ	7893	23	57	56	55

3.2 实验结果和分析

文[1](记为算法 B)是比较当前帧与所有已提取关键帧之间的差异程度,从而确定当前帧可否作为关键帧,这种方法的结果使得各个关键帧的内容之间最大限度的不同,从而更有代表性。为了更好地评价本文提出的方法,我们和该方法进行比较分析。图 2 是本文算法和算法 B 的对比实验结果,所使用的 352×240 原始图像取自《邂逅》中的一个视频序列共 3542 帧,图 2(a)是带宽 $h(h_c,h_m)=(18,0.25)$本文算法的检测结果,本算法提取了 14 帧关键帧,通过关键帧和字幕文字,我们能直观地和概括地了解这段视频发生了哪些事,有哪些人物。实验结果表明本文算法能够很好地利用帧间的时序关系和相似度来提取关键帧,从而达到较好的效果,其鲁棒性表现在大物体的快速运动上。图 2(b)是阈值为 3 的算法 B 的检测结果,虽然可以获得同本文算法比较接近的效果,但是算法 B 需要进行的阈值选取是点取值,它的微小差别对关键帧提取结果影响很大,阈值的选取难以把握,而本文算法采用的是均值漂移聚类方法,它所需设置的带宽 h 是范围取值,比算法 B 的阈值选取易于控制。



(a) 带宽$(h_c,h_m)=(18,0.25)$ (b) 阈值为3

图 2 关键帧提取的结果

这两种算法提取的关键帧都没有遗漏,但有少量冗余,这是由于光照较强的序列容易出现冗余帧。例如图 3 选用了 MTV 中的一个视频序列共 1879 帧,使用本文算法提取了 12 帧关键帧,显然最后一帧为冗余帧,原因在于舞台灯光的变化,其亮度变化很大,但是场景中人物的变化却由于时间短暂而变化很小(即舞台灯光变化前后的图像帧内容很相似),舞台灯光变化前后分别提取了关键帧,从而产生了 1 帧冗余帧。

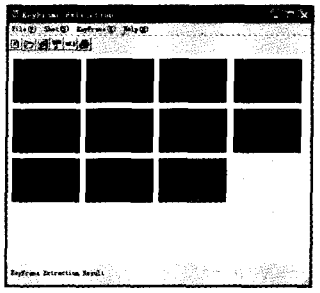


图 3 关键帧提取的结果

结论 本文提出了一种基于非参数密度估计聚类关键帧提取的算法,由于采用均值漂移聚类的方法,因此算法能够自动确定类别并且只需用户设置带宽 $h=\{h_c,h_m\}$即可。它综合利用了图像的颜色特征和运动特征,从而提高了关键帧提取的有效性。通过一些视频序列的提取实验和实验分析对比,我们认为,本文提出的提取算法对于无强光照的图像具有很好的提取效果,实验的提取结果与人的视觉具有良好的一致性。当然,该算法还有由于光照较强而产生冗余帧的问题,一个可能的途径是将图像的光照不均的校正技术融入算法中,可以使提取关键帧的有效性得到进一步的提高。

致谢 作者现在北京交通大学访问学习,此论文是在北京交通大学计算机学院须德教授和郎丛妍博士的指导下完成的,在此一并表示感谢!

参 考 文 献

1 Zhu X Q, Wu X D, Fan J P, et al. Exploring video content structure for hierarchical summarization. Multimedia Systems, 2004, 10(2): 98～115
2 Toklu C, Liou S P. Automatic keyframe selection for content-based video indexing and access. In: Proc. of SPIE. 2000, 3972: 554～563
3 Grabmeier J, Rudolph A. Techniques of cluster algorithms in data mining. Data Mining and Knowledge Discovery, 2002, 6(4): 303～360
4 Kim C, Hwang J N. An integrated scheme for object-bsed video abstraction. In: Proc. of the ACM Int. Conf. on Mutimedia. 2000. 303～311
5 Kin-Wai S, Kin-Man L, Guoping Q. A new key frame representation for video segment retrieval. IEEE Trans. On Circuits and Systems for VideoTechnology, 2005,15(9):1148～1155
6 Comaniciu D, Meer P. Mean Shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell, 2002,24(5): 603～619