

一种不需经验参数的视频镜头自校正聚类方法

熊 华

胡晓峰

(国防科学技术大学多媒体研究开发中心,长沙 410073) (国防大学模拟中心,北京 100091)

摘 要 镜头聚类是视频内容分析的重要途径.为能够自动、准确地实现镜头聚类,设计和实现了一种新的镜头聚类方法.这种方法从一个初始分割开始,经多次聚类分裂与合并的迭代,即能自动地进行误差校正,而且这种方法既不需要通过人工交互来解决试探聚类方法的误差调节问题,也不需要迭代聚类算法中进行难以确定的经验参数和经验阈值的设定.实验证明,该方法能较好地解决镜头的自动、准确聚类问题.

关键词 镜头聚类 合并与分裂 视频内容分析

中图法分类号:TN941.1 TN911.73A 文献标识码:A 文章编号:1006-8961(2001)03-0243-07

A Self-Adjusting Shot-Clustering Technique Without Experiential Parameters

XIONG Hua

(Multimedia R&D Center, National University of Defense Technology, Changsha 410073)

HU Xiao-feng

(Simulation Center of National Defense University, Beijing 100091)

Abstract Shot-clustering is an important issue in the field of video content analysis. The basic task of shot-clustering is to classify shots based on their low-level physical features. This paper describes a novel shot-clustering technique. Beginning with an initial classification of the shot set, our algorithm proceeds with merging and splitting iteration alternatively to reduce the errors in the initial results. The initial classification is based on the closeness of shots. The following clustering process is controlled by merging and splitting rules, which are based on the concepts of centroid, radius, intra-cluster hole, and inter-cluster space. The basic idea is that: (1) the intra-cluster hole of a cluster should be less than the intercluster space; and (2) the centroid distance of two clusters should be larger than the sum of their radiuses. Special consideration is put on the design of the iteration mode to suppress the possible errors. The main advantage of this algorithm is that it does not need any experiential parameters or thresholds, neither does it need any manual interaction, which is a basic requirement for automatic clustering of shots with no domain knowledge. Experimental results are presented and analyzed.

Keywords Shot-clustering, Merge and split, Video content analysis

0 引 言

随着多媒体时代^[1]的到来,视频已日益成为人们生活中不可或缺的信息载体,然而,由于视频媒体不像文本媒体那样便于检索和浏览,要在视频媒体中定位感兴趣的内容或快速掌握视频内容的梗概,

往往只能靠线性的快进、快退来实现,这就使视频信息的充分利用受到大大制约.为解决这一问题,近年已发展起来一种视频内容分析技术.所谓视频内容分析就是综合利用图象理解、模式识别、人工智能等领域的研究成果,来完成对视频信息的内容结构和内容标识的分析,为视频内容的检索、浏览和查询提供内容路标,它的应用,将为多媒体数据库管理与检

索、数字化图书馆、视频点播、电视节目编辑与管理、情报分析等,开辟一条崭新的道路。

镜头是视频序列编辑的最小单位,且多个镜头组织起来,就可对故事进行表达。由于每一个镜头包含一个事件或一组连续的动作,且每个镜头中的事件一定发生在一个场景中,而一个场景则可分散在多个镜头之中,另外一个故事线索也是分散在多个镜头之中,因此,表达更高层次的视频语义仅凭镜头是不合适的,需要在镜头之上做进一步的分析,以建立起镜头之间的语义逻辑关系,这就是所谓的镜头聚类(shot-clustering)。这种镜头聚类,即根据镜头的低层物理特征采用聚类方法对镜头进行分类的技术。由于不同的相似特征,将决定不同的关联内容,因此通过对镜头低层特征(例如颜色、纹理、声音等)的相似性分析,即可建立起镜头之间的内容关联。

目前的视频内容分析技术已达到镜头的准确探测和镜头特征的自动提取,但在镜头聚类分析的研究上还是刚刚起步。传统的聚类方法可分为层次聚类方法和分割聚类方法两种^[2],其中,分割聚类方法的输出是多个互不相交的聚类,而层次聚类方法的输出则是一个层次化的分类树。一般人们需要的是能输出多个分离聚类的分割聚类方法,而这种分割聚类方法包括试探算法和迭代算法两大类^[3],其中试探聚类算法是按照某种就近原则将被聚类样品逐一归入一定的聚类,虽然这类算法计算简单,不需经验阈值的设定,但由于没有聚类误差调节机制,其聚类效果常常难以预料,文献[4]中即采用这种方法对镜头进行聚类,并通过人工交互的方式来纠正试探算法的误差,而迭代聚类算法的基本思想是从一个初始分割开始,然后反复迭代直到某个控制函数达到最优为止,其中典型的迭代聚类算法有K均值算法^[2,5],ISODATA算法^[2,3]等,虽然通过多次迭代求精,这类算法能较好地解决试探聚类算法误差大的问题,但需要在聚类迭代运算开始之前就确定一些经验参数和阈值,如预期的聚类个数 K ,聚类的可容许均方差等,然而实际上由于视频文件可长可短,内容千奇百怪,往往很难对视频镜头的聚类情况作出正确的估计,因而 K 值等经验参数的确定也就无从入手。因此但若要达到自动、无领域知识的镜头聚类,则必须摒弃任何人工交互和经验参数、阈值。为此,本文设计和实现了一种新的镜头聚类算法。由于这种方法既不需要通过人工交互来解决试探聚类方法的误差调节问题,也

不需要进行迭代聚类算法中难以确定的经验参数和经验阈值的设定,因而较好地解决了镜头的自动、准确聚类问题。

1 镜头聚类过程描述

整个聚类算法从一个初始分割开始,经多次聚类分裂与合并的迭代,即能自适应地达到“最好的”聚类效果,且初始迭代的结果聚类集是后续聚类迭代的基础,初始迭代结束后,算法则按照合并与分裂的准则对初始聚类集进行交替迭代的合并、分裂处理,以修正初始迭代聚类结果的误差,直到合并和分裂迭代都不能再为止,图1即是该算法的整体流程。该算法的基本思路是通过迭代运算,以得到试探算法所达不到的精度,并通过合并判断和分裂判断规则的设定,来动态地从每次迭代后新产生的聚类集统计特性中来提取控制迭代收敛的阈值、参数,从而避开了经验参数或阈值的预先设定。

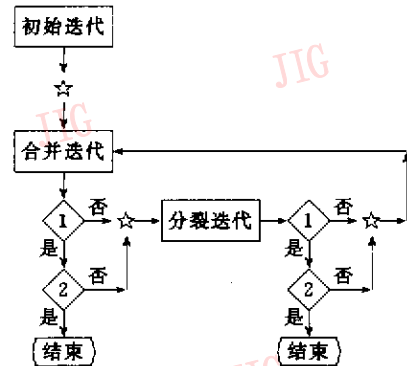


图1 镜头聚类算法的整体流程

说明:

判断1:判断本次迭代实际上是否未做任何操作;

判断2:判断前次迭代实际上是否未做任何操作,当连续两次迭代未执行任何操作时迭代终止。

☆表示各次迭代的输出结果聚类集,也是紧跟其后迭代的输入。

1.1 初始迭代

这里,初始迭代的任务是对被聚类镜头集进行一个大致的划分,并尽量把相似的镜头放在同一个聚类中,以便为后续的分裂合并微调打好基础。

该初始迭代的唯一输入是距离矩阵,而距离矩阵的元素描述的则是待聚类样品镜头集中两两镜头之间的距离,同时这个距离可由镜头的任何低层特征或低层特征组合的相似性测度来计算得到,且距离越近的两个镜头(按照某种低层特征)越相似。若镜头 S_i, S_j 的相似度为 $M(S_i, S_j)$,则它们之间的

距离 $D(S_i, S_j)$ 为

$$d_{ij} = D(S_i, S_j) = 1 - M(S_i, S_j) \quad (1)$$

而且该距离矩阵是个对称矩阵.

初始迭代算法的基本原则是距离最近的镜头最有资格被聚为一类,依据这一原则,初始迭代算法即按照距离矩阵中最小距离排队,依次聚类相应的镜头.它首先将距离最近的两个镜头聚为一类,然后将次近的镜头聚为一类,依此类推,直到所有镜头都有所归属,而在次小距离(大于等于当前最小距离的最小距离)搜索过程中,若构成次小距离的两个镜头之一已被聚入某类 A ,则将另一镜头也聚入类 A ,不另外成立新类.

1.2 合并迭代

合并迭代的任务就是将那些在初始迭代聚类结果中,分得过小、过碎的聚类对进行合并,为此需首先引入“聚类质心”、“聚类半径”和“类间空隙”的概念.这里,所谓“聚类质心”是指一个聚类中的某一个镜头,若它满足该类中其他所有镜头离该镜头距离之和最小,则该镜头即为聚类质心;“聚类半径”则是指离质心最远的镜头距质心镜头的距离;“类间空隙”是指两个聚类所含元素间的最近距离.

满足如下条件的聚类对可考虑合并:①两个聚类质心的距离小于两个聚类半径之和;②聚类对各聚类的半径小于整个聚类集的平均半径,其中,第1个条件是借用几何学上关于两圆相交的定义,第2个条件是为了抵消噪声点对合并判断的影响,另外,被合并聚类中,若包含单镜头聚类,则该单镜头不能是最近才从候选聚类对中的另外一个非单镜头聚类中剔除的镜头,这样做是为了防止同一个镜头不停地被相邻的合并、分裂迭代所操作,而产生所谓的“振荡”效应,故只对下述两对候选聚类进行上述合并判断:其一是聚类集中质心相距最近的聚类对,这是按照聚类合并准则最有资格合并的聚类对;其二是聚类集中类间空隙最小的聚类对,把这一聚类对作为候选聚类对是为了加速分裂迭代的收敛速度和减轻噪声对分裂判断的影响.上述候选聚类对,若通过合并判断,则可合并之,并记录此次合并操作所移动的镜头个数,以指导随后的分裂迭代,否则,若合并判断失败的原因是某个聚类的半径大于、等于聚类集的平均半径,则标记该聚类为待分裂聚类,在随后的分裂迭代中进行分裂判断和操作,因为这种情况可能指示该聚类含有噪声镜头.图2为合并迭代的流程图,其中,合并判断与处理模块的流程图如图3.

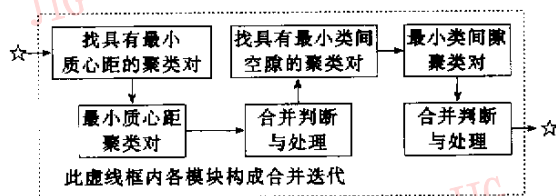


图2 合并迭代流程图

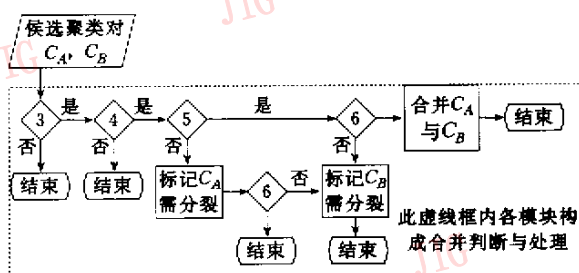


图3 合并判断与处理流程图

说明:

判断3: C_A, C_B 的质心距是否小于 C_A, C_B 的半径;

判断4: 单镜头聚类是否为刚被剔除的元素;

判断5: C_A 的半径是否小于聚类集的平均半径;

判断6: C_B 的半径是否小于聚类集的平均半径.

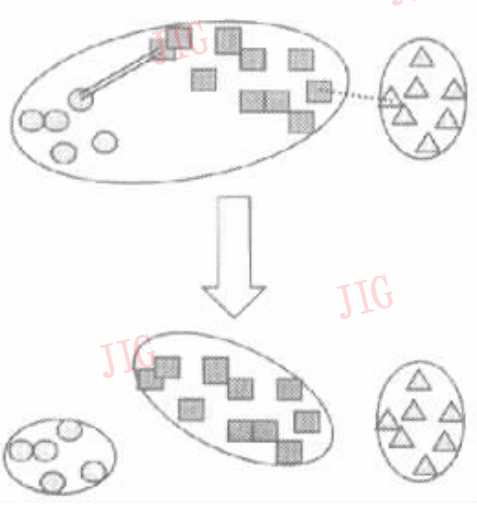
1.3 分裂迭代

由于在任何被聚类镜头集中都会存在一些零散镜头,它们不应该归于任何类,因此在初始迭代算法中,这些零散镜头则按照就近原则被硬性分到各个聚类中,从而使得这些镜头成为其所在聚类的“噪声点”,而分裂迭代的任务就是通过聚类的分裂操作,将初始迭代聚类结果中的这类噪声点剔除.为了方便叙述,首先引入“最大类内空洞”的概念.大家知道,任何一个聚类都可看作由两个或多个子类构成,而“类内空洞”就是指这些子类之间的空隙,且各子类之间的空隙将构成一个对称矩阵,而“最大类内空洞”则是这个矩阵中的最大值.

当某聚类的最大类内空洞大于、等于整个聚类集的最小类间空隙时,就可以考虑分裂该聚类(见图4).

这种最大类内空洞的计算可以用该类的“最大最小距离”来近似,一般聚类中某元素的“最小距离”是指该元素距其所属类中的其他元素的最短距离,某聚类的“最大最小距离”则是指该聚类中所有镜头的最小距离中的最大者,这种持有最大最小距离的元素也是该类可能的噪声点,而分裂操作的实质则是将通过分裂判断,将聚类中的最大最小距离持有者从该类中剔除,即一次分裂操作只移动一个镜头.另外,由于过小的类间空隙可能指示着需合

并的两类,所以应尽快将具有最小类间空隙的聚类对合并(见上节),以提高分裂“门槛”的精度。

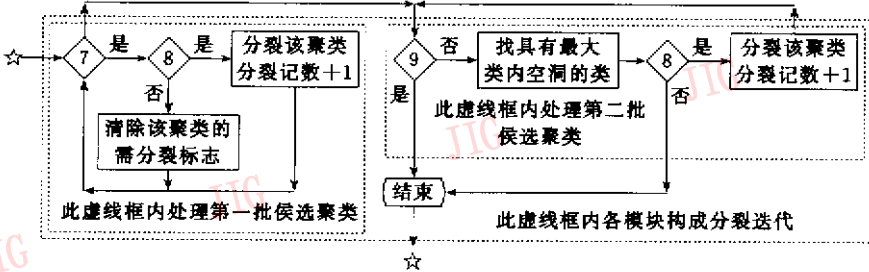


箭头上面为分裂前的情况,下面为分裂后的情况。由圆点和方点构成的椭圆为应分裂的类。双实线表示类内空洞,虚线表示类间空隙。由于类内空洞大于类间空隙,应该执行分裂操作。

图4 应分裂的类

这里若只对两种候选聚类进行分裂判断,则首先选择前次合并迭代中所标记的需分裂的聚类,之所以要优先考虑这种聚类,是为了淡化噪声对合并判断中半径约束的影响;其次选择那些类内空洞最大的聚类(因为这是按照分裂准则最有资格被分裂的聚类)。

对上述第1批候选聚类,之所以要全部进行分裂判断,并对实际操作镜头数计数,是为了确保在合并判断中消除可能的噪声;而对第2批候选聚类,若满足下列条件之一,即终止此次分裂迭代,并进入下次合并迭代:①若当前具有最大类内空洞的聚类不能通过分裂判断,则表明当前聚类集中所有聚类都不再需要分裂;②若当前已操作镜头数超出本次迭代可操作镜头数的上限,即当前次合并迭代操作镜头数为0时,则可操作镜头数上限为1;否则,可操作镜头数上限即等于前次合并迭代所操作的镜头数,这是为了使相邻两次合并、分裂迭代的操作幅度相当,以便互相帮助,来逐步抑制噪声对方迭代的影响。其分裂迭代的流程图如图5。



说明:
判断7:是否存在标记为需分裂的类;
判断8:该聚类的最大类内空洞是否大于最小类间空隙;
判断9:分裂计数是否超过此次分裂迭代的上限。

图5 分裂迭代

1.4 算法收敛性讨论

如前所述,当合并和分裂迭代都不能再进行时,则整个算法终止,即算法终止的条件是整个聚类集中再没有任何聚类能通过分裂判断和合并判断。一般来说,随着分裂、合并迭代的进行,能通过合并、分裂判断的聚类会越来越少,算法会自然收敛,但是,有时也会产生相邻的合并、分裂迭代操作同一组镜头,这样将使得这些镜头刚被分裂出去,在下次合并迭代中又被合并回来,从而导致迭代过程无法收敛,这种现象即称为“振荡”,而且这种振荡可能是导致算法不收敛的唯一原因。根据产生振荡的镜头数目,振荡可分为如下两种:其一是“简单振荡”,它是指单个镜头刚从某聚类中被剔除,在紧接其后的合并迭代中又作为单镜头聚类被合并回原聚类,但是合并判断的第3个条件就杜绝了这种振荡的发生(见1.2节);其二是“复杂振荡”,它是指某聚类中的一组镜头被从该

聚类中剔除后,又合成一个新的聚类,然后这个新合成的多元素聚类又被合并回原聚类,从而造成改组镜头与原聚类构成的振荡,但这种振荡是不可能发生的,因为按照分裂操作的设计,所有刚被剔除的镜头都是作为单元元素聚类存在的,而按照合并判断,该两个单元元素聚类是不能被合并的,这是因为单镜头聚类的半径为0,两个单镜头聚类的半径和仍为0,从而它们的质心距不可能小于半径和,从而更不可能发生由被剔除的一组镜头合成的新聚类杀回原聚类的情况,因此本文的镜头聚类迭代过程总会收敛。

2 实验结果与分析

本文共对近580min的视频文件进行了基于上述算法的镜头聚类实验,其涉及的视频种类包括教学材料、新闻、故事片,这种视频的长度为2min-

40min 不等.实验证明,虽然本文的算法未使用任何人工交互和经验参数,但通过合并、分裂迭代,初始迭代的聚类效果仍能得到自动的校正,且主观视觉效果非常好.这个算法对长视频的效果比对短视频的效果好,这可能是由于被聚类样品集越多,其统计特性就越明显的缘故.另外,由于聚类算法的唯一输入是距离矩阵,因而通过灵活采用不同低层特征来计算距离矩阵,即可得到各种不同意义的聚类结果.本文在实验中,分别采用了基于镜头运动特征^[2]的距离矩阵和基于镜头颜色特征^[6]的距离矩阵来测试本文的聚类算法.通过实验发现,用前一种距离矩阵进行聚类,可把以相似运动方式变化的镜头聚在一起,而以后一种距离矩阵进行聚类的结果,则能指示包含同一对象的镜头或发生在同一地理位置的镜头.如果将这些聚类结果善加利用,对视频内容浏览和视频内容检索都会大有裨益.

由于篇幅的缘故,这里只给出一段较短视频的部分实验结果.

素材为一段长 263s 的 MPEG-1 文件.实验时,首先对该视频进行镜头分割,即先使用镜头探测软件对该视频进行自动分割,然后再进行人工调整,以得到基本上无漏判和误判的视频文件镜头集,共 40 个镜头,其代表帧取的是视频流中离镜头中间帧最近的 1 帧^[7].此实验采用基于代表帧颜色特征^[6]的距离矩阵进行聚类.

通过初始迭代共分出 12 个聚类,其各聚类质心镜头对应的代表帧按聚类在初始迭代运算中的生成先后排序如图 6.图 7 列出了部分聚类所含镜头的代表帧,每一行图片是一个聚类,各代表帧按距质心镜头远近排序,行尾是该聚类的编号.由图 7 可见,初始迭代已有较好的效果,如视觉上相似的镜头已被聚在一起,不过从图 6 和图 7 也可以较明显的看出,有些聚类尚需要合并,如聚类 1 和聚类 2,聚类 3 和聚类 4.从图 7 也能看出有些聚类含有明显的噪声点,尚需要分裂,如聚类 4 的最后一个镜头,聚类 7 的最后一个镜头等.



图 6 初始迭代实验结果(各聚类质心一览,按类生成先后排序)



图 7 初始迭代结果(聚类 1 2 3 4 7 所含镜头代表帧)

该实验共迭代运算 15 次,最后共得到 9 个聚类和 1 个杂类,其与各聚类质心镜头对应的代表帧见图 8,图 9 则列出了部分聚类所含镜头的代表帧.各次迭代的实验数据列于表 1.将图 6,图 7 与图 8,图 9 进行比较可以看出,通过本文算法迭代运算 15 次后,初始迭代结果中的聚类 1 和聚类 2,以及聚类 3 和聚类 4 被正确地合并在一起,且聚类 4、聚类 7 中的噪声点也被剔除,初始迭代中的误差也得到了较好的校正.由表 1 可以看出分裂判断的依据——最小类间空隙经过第 1 次合并迭代后,就不再变化了,且合并迭代在一开始就把

分裂迭代所受的误差影响降到了最低,这说明把有约束的合并迭代放在分裂迭代之前是一种较好的做法.随着迭代的进行,整个聚类集的平均半径不断减小.另外,从表 1 中还可看出,合并迭代的效率较高,整个算法大部分时间都在等待分裂操作终止,而且本文的算法在设计分裂操作个数上限时,也反映了这一规律.从图 9 还可以看出,镜头聚类结果在一定程度上还反映了语义内容,比如聚类 1 是关于街头小贩的镜头,聚类 2、5 则是关于男主角分别在不同地方的镜头.



图 8 最终实验结果(各聚类质心一览,按类生成先后排序)

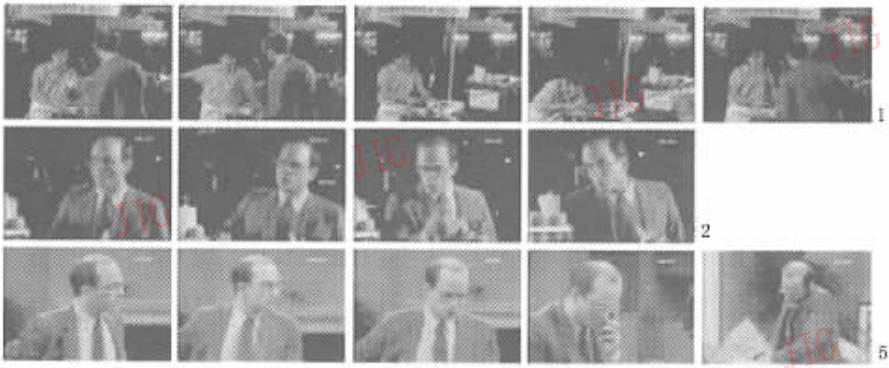


图 9 最终迭代结果(图中显示聚类 1 2 5 所含镜头代表帧)

表 1 各次迭代的数据

序号	聚类个数	类型	空迭代	最小质心距	最大类内空洞	最小类间空隙	被操作镜头数	平均半径
1	12+0	初始	否	0.130 421 5	0.460 191 7	0.101 313 9		0.271 274 2
2	10+0	合并	否	0.222 182 6	0.460 191 7	0.215 293 6	4	0.312 531 9
3	10+4	分裂	否	0.222 182 6	0.274 064 9	0.215 293 6	4	0.203 557 0
4	10+4	合并	是	0.222 182 6	0.274 064 9	0.215 293 6	0	0.203 557 0
5	10+5	分裂	否	0.257 492 8	0.274 064 9	0.215 293 6	1	0.189 985 8
6	10+5	合并	是	0.257 492 8	0.274 064 9	0.215 293 6	0	0.189 985 8
7	10+6	分裂	否	0.257 492 8	0.272 893 0	0.215 293 6	1	0.180 643 6
8	10+6	合并	是	0.257 492 8	0.272 893 0	0.215 293 6	0	0.180 643 6
9	10+7	分裂	否	0.257 492 8	0.231 545 9	0.215 293 6	1	0.173 928 1
10	10+7	合并	是	0.257 492 8	0.231 545 9	0.215 293 6	0	0.173 928 1
11	10+8	分裂	否	0.257 492 8	0.228 456 4	0.215 293 6	1	0.166 474 6
12	10+8	合并	是	0.257 492 8	0.228 456 4	0.215 293 6	0	0.166 474 6
13	10+9	分裂	否	0.228 456 4	0.212 346 0	0.215 293 6	1	0.162 899 8
14	10+9	合并	是	0.228 456 4	0.212 346 0	0.215 293 6	0	0.162 899 8
15	10+9	分裂	是	0.228 456 4	0.212 346 0	0.215 293 6	0	0.162 899 8

注:聚类个数一栏中“+”之后的数字表示被剔除的噪声点数.空迭代一栏中“是”表示此次迭代未执行任何实质性操作;“否”则反之.

3 结 论

为了避免进行难以确定的经验参数和阈值的设定,本文设计了一种新的自校正镜头聚类方法.这种方法先用初始迭代算法对整个镜头集作大致分割,再用特别设计的合并和分裂算法对初始聚类结果作交替迭代的微调.从实验结果看,这种镜头聚类方法能够达到(主观感受上)较为准确和合理的聚类自校正效果.为使镜头聚类真正在视频内容分析中发挥作用,还需就以下两个方面作进一步研究:其一是研究低漏判率的镜头自动探测方法;其二是在视频内容分析中如何应用镜头聚类.由于前者是实现镜头自动聚类的基础,目前所采用的镜头探测算法^[7]还不够准确,漏判率较高,因此需要改进;而后者主要包括研究如何利用镜头聚类,以及如何为镜头数据库建立索引^[7]和如何利用镜头聚类来构造视频内容的层次结构(如场景、故事单元等).

参 考 文 献

- 1 胡晓峰,李国辉.多媒体系统.北京:人民邮电出版社,1997:8.
- 2 Zhong D,Zhang H J,Chang S-F. Clustering methods for video browsing and annotation. In: Proceedings of SPIE Conference on storage and retrieval for image and video database: IV, San Jose CA USA, 1996: 239246.

- 3 沈清,汤霖.模式识别导论.长沙:国防科大出版社,1990:8.
- 4 Yeung M,Yeo B-L,Wolf W *et al.* Video browsing using clustering and scene transitions on compressed sequences. In: Proceedings of SPIE Conference on multimedia computing and networking, Bellingham WA USA, 1995: 399413.
- 5 薛峰.基于内容检索的图象和视频存储结构和索引技术的研究和实现[硕士学位论文].长沙:国防科学技术大学七系,1999.
- 6 柳伟.基于内容的图象检索技术研究和实现[硕士学位论文].长沙:国防科学技术大学七系,1999.
- 7 曹莉华.视频媒体的基于内容处理和检索的研究与实现[博士学位论文].长沙:国防科学技术大学七系,1998.

熊 华 国防科学技术大学博士研究生.主要研究方向为视频内容分析、视频的基于内容检索与浏览.