专题报告

视频分层组织方案和技术

章毓晋，陆海斌
（清华大学电子工程系，北京 100084）

[摘要] 数字视频是多媒体信息系统中重要的数据类型。传统的视频表示方法仅是一个时间序列——媒体流，所以对计算机来说，很难在内容层次上认知视频。为了有效地访问和利用视频信息，合适的视频数据组织非常重要。文章提出将视频划分成四个层次即视频节目、情节、镜头和图像帧的组织方法。这样一种分层结构提供了紧凑和有意义的视频目录，方便了视频非线性浏览和基于内容的检索。为了得到这样一种组织，不仅要检测出镜头和情节这些视频单元的边界，还要提取镜头关键帧和选择情节有代表性的镜头和代表帧。文章介绍一系列分割视频和组织视频的准则和方法，并把它们结合起来组成了一个原型实验系统。文中还给出了一些对实际视频进行组织的结果，它们表明该组织形式是非常有效的。

[关键词] 视频组织；浏览；镜头；关键帧；情节；代表帧

1 前言

数字视频是多媒体信息系统中重要的数据类型，其特点是数据量大、信息量也大[1]。传统的视频表示方法是将视频表示为比特序列——视频流，所以要利用视频的内容进行索引、浏览、查询、检索等就需要对视频进行有效合理的组织。图 1 给出视频信息组织和其它应用的联系。这里，采集的视频存于一个视频数据库中。为了让视频数据库中的原始视频数据可以通过网络浏览或检索，需要对它们进行分析、索引和组织。组织后的视频数据具有合适的结构，适用于非线性浏览和基于内容的检索。

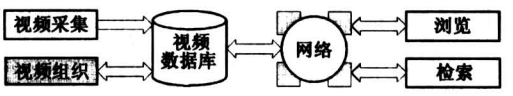

图 1 视频组织和应用示意图

Fig.1 Outline of video organization and application

本文提出一种视频组织方案，它基于镜头和情节这样的视频单元。第 2 节给出整体组织方案的工作框架和结构，第 3 节和第 4 节分别介绍镜头检测和关键帧提取、情节确定和代表帧选择的技术。基于对视频单元的紧凑表达，第 5 节给出分层组织的方法和实际视频组织的示例。

2 组织方案

用于组织的分层结构（方案）将视频分成四层，即视频节目层、情节层、镜头层和帧图像层，如图 2。根据这个方案可对视频进行三类操作：组织、浏览和检索。

1）视频组织将视频元素/组元按照某种（事先确定的）结构联系起来，以提供对快速浏览和检索的支持。视频组织从最低一层——帧图像层开始。这层对应原始视频数据，即时间序列图像帧。借助镜头检测，可将图像帧组合/聚合成镜头。镜头是视频的一种基本单元，它包括按时序连接的一组帧图像，各个镜头在相同的场景拍摄，包含空间中某个位置的一个连续动作。借助一些高层知识，可将

[收稿日期] 1999－12－20
[基金项目] 国家自然科学基金资助项目（69672029）和国家高技术研究发展计划资助项目（863－317－9604－05）
[作者简介] 章毓晋（1954－），男，江苏江阴市人，清华大学教授，博士生导师

一些镜头（不一定相连或相邻）结合成情节。情节是视频的一种语义单元，它一般描述一段故事或行动。换句话说，情节中的镜头内容上是相关的，但可以在时序上是不连续的或空间上是分离的。每个视频节目（如电影）由一系列情节构成。由以上讨论可知，视频组织是对视频流不断抽象的过程。

2）视频浏览是指用户在视频数据中“航行”(navigate)，并发现感兴趣的视频片段/序列。为了对视频数据库进行浏览，最好有个全面的概述（目录），而对视频节目，一个合适的概要也能起到这样的作用。这样一个概要不仅要提供节目的主要内容，而且要提供进入节目不同部分的多个入口。一旦将一个视频节目组织成上面所讨论的结构，快速浏览就变得很直接、方便。浏览可从节目层开始由上向下进行。浏览者首先可进入情节层，发现感兴趣的情节。因为每个情节包含若干个镜头，所以进入情节后再浏览有关的镜头是很容易的。镜头由一序列图像帧组成，找到感兴趣的镜头后，浏览各图像帧是很直接的。

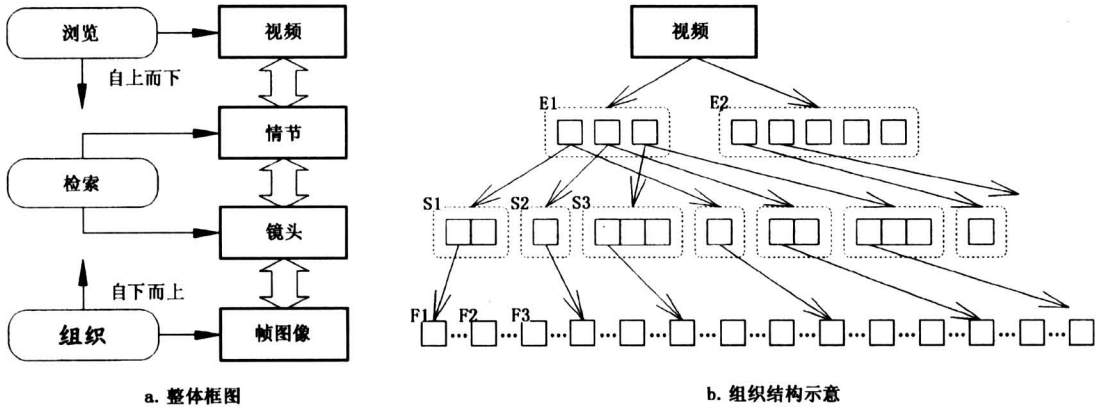

图 2　视频组织框架

Fig.2　Framework for video organization

3）视频检索是一种直接寻找感兴趣的视频序列的方法或过程。与由下向上的组织和由上向下的浏览不同，基于内容的视频检索既可在镜头层进行、也可在情节层进行。在镜头层，每个镜头可用其关键帧表示。在情节层，每个情节可用其代表帧表示。这样，在镜头层的检索可借助镜头的关键帧进行，在情节层的检索可借助情节的代表帧进行。在这两种情况下，因为检索对象对应单帧图像，所以可使用各种基于内容对静止图像检索的具体技术来进行。另外对镜头关键帧和情节代表帧进行文字注解后，现有的文字检索技术也可结合进来。

3　镜头检测和关键帧提取

镜头是视频的一种基本单元，它由时间上相连的一组帧图像组成。镜头检测是要将视频流切成一个个分离的镜头。这时需要确定镜头的时间边界，或者说要检测镜头的转变或切换处。常见视频节目中的镜头切换可分两种[2]：一种是直接的（突然的）切换，称为切变；另一种是光学切换，是对应场景的逐渐变化，称为渐变。前者切换是在两帧图像间进行的，两帧图像的视觉感受可以完全不同；而后者的变化一般跨越若干帧图像，从视觉上看镜头转换比较平滑和舒适。检测这两种切换的一种策略是顺序检测它们：先检测切变，后检测渐变。

切变检测的工作流程和主要步骤见图 3。输入的视频流既可以是原始的视频流，也可以是压缩后的视频流。对前者利用邻域平均，对后者提取直流分量，都可得到待检测的视频流。这种预处理可减少检测算法对摄像机和物体运动的敏感性，并同时滤去视频采集带来的噪声。视频节目在拍摄时常利用闪光，它会造成场面光强的较大变化而使得切变算法误认为是镜头切变。所以，还要将闪光位置检测出来，并从可能的切变位置中除去，以得到从真正切变位置分开的镜头序列[3]。

在切变检测的基础上可进一步进行渐变检测。影片中常用的渐变主要有淡入、淡出和叠化。这些

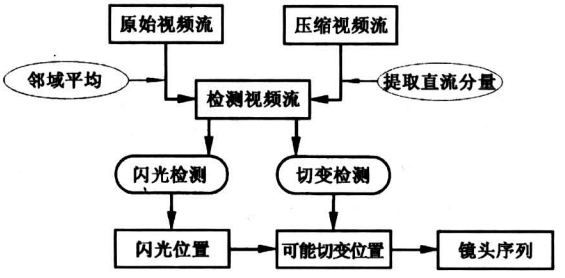

图3 切变检测流程图

Fig.3 Diagram for cut detection

渐变均可用一种视频模型来统一描述[4]。根据这种模型，可将渐变位置的检测转化为求取帧图像方差序列的极值问题。根据几十分钟视频的检测实验，对淡入和淡出的检测均可达到接近100%的查全率和准确率[2]。而对叠化的检测，目前检测的查全率和准确率也能平均达到90%左右。表1给出对含有19个叠化的2个视频片段的实验结果。

表1 叠化检测实验结果

Table 1 Results of dissolve detection

视频	总数	正确	丢失	误判	查全率	准确率
"Casablanca"	10	9	1	1	90%	81%
"CONAN"	9	9	0	1	100%	90%

将各镜头检测出来后，对每个镜头可提取关键帧，并用关键帧简洁地表达镜头。这是因为每个镜头都是在同一个场景下拍摄的，同一个镜头中的各帧图像有相当的重复信息。镜头的关键帧就是反映该镜头中主要信息内容的帧图像，一般一个镜头要用所提取出的一个或若干个帧图像来表示。另外，用关键帧表示镜头，使得对视频镜头可用基于图像的技术进行检索。

实际应用时，由于场景中目标的运动或拍摄时摄像机本身的操作（如变焦，摇镜头等），一个镜头用一幅关键帧来表示不太够。关键帧应能提供一个镜头的全面概要，或者说应能提供一个内容尽量丰富的概要。从这个角度说，关键帧的提取可看作一个优化过程。根据信息论的观点，不同（或相关性较小）的帧图像比类似的帧图像携带较多的信息。所以当需要提取多幅关键帧时，用于关键帧提取的准则是考虑它们之间的不相似性，根据各镜头的不同特点，用下列方法自动地提取1～3帧图像。

设用 f 表示一帧图像，$S=\{f_n, n=1, 2, \cdots, N\}$ 表示具有 N 帧的一个镜头，取帧图像 f_1、$f_{N/2}$和 f_N 作为候选关键帧。先定义两幅图像 f_i 和 f_j 间的差异距离为：

$$D(f_i, f_j)=\sum_{x,y}|f_i(x,y)-f_j(x,y)|。(1)$$

提取镜头关键帧时先计算两两候选关键帧之间的距离，即 $D(f_1, f_{N/2})$、$D(f_1, f_N)$ 和 $D(f_{N/2}, f_N)$，并将它们与一个预定的阈值 T 比较，按下列准则确定关键帧：

1）如果它们都比 T 小，说明它们之间比较接近，此时取 $f_{N/2}$作为关键帧；

2）如果它们都比 T 大，说明它们之间差距较大，需要将它们都取为关键帧；

3）在其它情况下，取距离最大的两帧图像作为关键帧。

4 情节划分和代表帧选择

情节是视频的一个语义单元，它一般包含描述一个故事的一个至多个镜头。目前还没有完全自动将视频分成语义级情节的方法。一种半自动检测逻辑故事单元（logical story unit）的方法可近似检测出情节[5]。它利用自动技术对视频进行预处理，然后借助用户交互来划分情节。

如何有效地表达情节也是很重要的。由于情节常包含多个镜头，所以要从这些镜头中进行选择。直观的方法是选择最重要的镜头，但仅用一个镜头代表一个情节容易丢失其它镜头的信息，而且1个完整镜头的数据量仍很大。由于前面已将镜头关键帧提取了出来，所以可从情节中所包含镜头的关键帧集合中来选择代表该情节的代表帧。考虑到视频拍摄和人们观看的习惯和特点，以下两条准则在选择情节代表帧时需要注意：

1）反复出现的镜头比较重要；

2）延续时间较长的镜头比较重要。

根据这两条准则，我们设计了一种新的聚类方法来选取情节代表帧，它包括以下几个步骤：

1）分层聚类 设一个情节中有 N 个镜头，从该情节中先选取第一个镜头 S_1 的关键帧 F_1（这里设每个镜头只用一幅关键帧），建立一个类 C_1，把 F_1 放入类 C_1，并计算 F_1 与该情节中其它镜头关键帧的距离 $D(F_1, F_i)$，$i=2, 3, \cdots, N$。如果这个距离小于预定的阈值 T，就将两镜头聚在同一类，否则建一个新类。上述过程依次对情节中

的每个镜头进行，最后可将情节中的所有镜头划到一定数量的类中。

2）模糊分类　设在第一个步骤中最后得到 K 个类，对每个类别 C_i 和镜头 S_j 计算下列模糊隶属度函数：

$$\mu_{ij} = \sum_{k=1}^{K} \left(\frac{D(S_j, E_i)}{D(S_j, E_k)} \right)^{\frac{-2}{m-1}}, \qquad (2)$$

其中 m 是模糊指数，E_i 是类别 C_i 的中心。E_i 的作用是根据上述第一条准则对反复出现的镜头进行加权。另外根据上述第二条准则对模糊隶属度函数用镜头时间长度进行加权。根据加权结果，最后从情节中选出各类中模糊隶属度函数值最大的镜头，将它们的关键帧组合成该情节的代表帧集合。

5　实验结果和讨论

上述技术已被集成在一个试验系统中。这个系统的功能包括：读取视频，检测镜头，提取镜头关键帧，确定情节，选择情节代表帧，注解情节和镜头。一段视频经过上述处理和组织后，可建立起如图 2 所示的结构，此时就可以方便地进行浏览和检索了。

下面具体介绍和讨论一组试验结果，其中使用了一段 54 秒长的电影片断（源自“空军一号”）。利用第 3 节介绍的技术，将这段视频片断分成 23 个镜头。又利用第 4 节介绍的技术，将这些镜头组合成 2 个情节。组合后的结果见图 4。在图 4 中，第一行是对应这两个情节的两组情节代表帧，第二行是对应第一个情节的八个关键帧。图中的箭头指示了第一个情节中各代表帧和关键帧的联系。用户从第一行选择情节和代表帧，就很容易进一步观看其关键帧在第二行给出的感兴趣镜头。

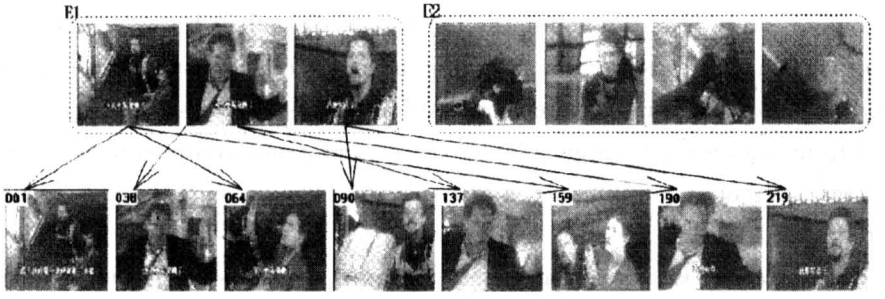

图 4　组织结果示例

Fig.4　Experimental organization results

对视频的组织不仅提供了快速浏览和检索的基础，也提供了一种对视频数据紧凑的表达形式。以上述实验的数据为例，其组织结果分四层归纳在表 2 中。由表 2 可见，当把视频流中的 1 350 帧图像用对应 23 个镜头的 60 幅关键帧表达后，已经取得了相当于 22∶1 的数据压缩比。进一步将这些镜头组合成 2 个情节，并用 7 幅代表帧表达，相应的压缩比增加到约 200∶1。

表 2　对“空军一号”中 54 秒视频的组织结果

Table 2　Organization results for a videoclip of 54 s in “Airforce No. 1”

层	单元数	表达	压缩比
(1) 视频	1 段		
(2) 情节	2 情节	7 代表帧	≈200∶1
(3) 镜头	23 镜头	60 关键帧	≈22∶1
(4) 视频流	1 350 帧	1 350 帧	1∶1

6　结语

文中提出了一个新的视频组织方案，并根据这个组织框架，通过结合所设计的镜头检测和镜头关键帧提取方法，以及情节划分和情节代表帧选择方法，构建了一个视频组织实验系统。对真实视频的组织试验结果是令人满意的。

进一步的工作包括结合基于视觉线索和文字搜索的检索，以及结合由下向上的组织结构和由上向下的知识驱动，以提高视频组织系统的性能。

参考文献

[1] Nwosu K C, Thuraisingham B, Berra P B. Multimedia database systems—a new frontier [J]. Multimedia, 1997, (3): 21～23

[2] Patel N V, Sethi I S. Video shot detection and char-

acterization for video databases [J]. Pattern Recognition, 1997, 30: 583～592

[3] Lu H B, Zhang Y J. Detecting abrupt scene change using neural network [A]. Proceedings of Visual Information Systems'99 [C], 1999. 291～298

[4] Lu H B, Zhang Y J, Yao Y R. Robust gradual scene change detection [A]. Proceedings of ICIP'99 [C], 1999

[5] Hanjalic A, Lagendijk R L, Biemond J. Automated segmentation of movies into logical story units [A]. Proceedings of Visual Information Systems'99, 1999 [C], 229～236.

Scheme and Techniques for Hierarchical Oranization of Video

Zhang Yujin, Lu Haibin

(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

[**Abstract**] Digital video is an important data format in multimedia information systems. Traditional video representation is just a time sequence —— video stream, thus it is difficult for computer to recognize or perceive video in the content level. To efficiently access and utilize video information, a suitable organization of video data is critical. This paper proposes a video organization scheme, which arranges video into four layers: video program, episode, shot and image frame. This hierarchical structure provides a compact and meaningful video catalogue, which can be easily used for non-linear browsing and content-based retrieval of video data. To achieve such an organization, it needs not only detect the boundary of shots and episodes, but also extract the key frames of shots and select the representative shots and frames for episodes. This paper proposes a number of suitable criteria and techniques for video segmentation and organization, and integrates these techniques into a prototype system. Some organization results using real video data are presented, which show the effectiveness of this organization scheme.

[**Key words**] video; organization; browsing; shot; episode; key frame; representative frame

作者·编者

敬告作者

为适应我国信息化建设需要，扩大作者的学术交流渠道，本刊已加入《中国学术期刊（光盘版）》和中国期刊网，其相应的作者文章著作权使用费交中国版权保护中心统一处理。如作者不同意将文章编入上述数据库，请在来稿时声明，本刊将作适当处理。

《中国工程科学》
编辑部