一种层次的电影视频摘要生成方法

赵亚琴[1] 周献中[2] 何 新[1]
[1](南京理工大学自动化学院,南京 210094) [2](南京大学工程管理学院,南京 210093)

摘 要 合理地组织视频数据对于基于内容的视频分析和检索有着重要的意义。提出了一种基于运动注意力模型的电影视频摘要生成方法。首先给出了一种基于滑动镜头窗的聚类算法将相似的镜头组织成为镜头类;然后根据电影视频场景内容的发展模式,在定义两个镜头类的 3 种时序关系的基础上,提出了一种基于镜头类之间的时空约束关系的场景检测方法;最后利用运动注意力模型选择场景中的重要镜头和代表帧,由选择的代表帧集合和重要镜头的关键帧集合建立层次视频摘要(场景级和镜头级)。该方法较全面地涵盖了视频内容,又突出了视频中的重要内容,能够很好地应用于电影视频的快速浏览和检索。

关键词 电影视频 场景检测 摘要生成 运动注意力模型
中图法分类号: TP391 TN941.1 **文献标识码**: A **文章编号**: 1006-8961(2007)08-1412-06

Automatically Generating Hierarchical Summary for Film Video

ZHAO Ya-qin[1], ZHOU Xian-zhong[2], HE Xin[1]
[1](*School of Automation, Nanjing University of Science & Technology, Nanjing* 210094)
[2](*School of Management and Engineering, Nanjing University, Nanjing* 210093)

Abstract It is important to properly organize the unstructured video data for content-based video analysis and retrieval. In this paper, we propose a unified approach for film video summarization based on the analysis of video structure and motion attention model. Video shots are firstly grouped into shot clusters. Afterwards, according to the characterization of film video scene, a temporally and spatially integrated strategy is presented to parse shot clusters into semantic scenes in terms of the definition of temporal relationships between two shot clusters. Finally, representative frames and highlight shots are selected from scenes by using motion attention model. The scheme offers an efficient mean for browsing and effectively retrieving film video.

Keywords film video, scene detection, summarization generation, motion attention model

1 引 言

视频摘要的自动生成是目前基于内容的视频检索的研究热点之一,视频摘要主要分为两类:视频概要(video summary)和缩略视频(video skimming)。视频概要是从原始视频中剪取或生成的一小部分静止图像的集合。而缩略视频保持了视频内容随时间动态变化的视频固有特征,由一些浓缩视频内容的视频片段(镜头)组成。视频概要的建立过程快速简单,并且一旦建立就可以很方便地显示和组织,并支持视频快速导航。因此,目前关于视频摘要的多数研究主要集中在静态视频摘要上[1~4]。

视频概要生成主要有基于镜头的关键帧提取方法[5~7]和基于场景的代表帧选择方法[8,9]。镜头的关键帧提取主要有 3 类:基于图像信息的方法[5]、基于运动的方法[6]和全景图拼接法[7]。但对于一个长的视频文档,镜头的关键帧数量是相当可观的,这样浏览起来不仅费时,而且效率很低。基于场景的代表帧选择可以用更少量的关键帧表达视频,常用的方法是视频帧聚类法[8,9],但需要解决聚类数目如何确定,场景的重要程度如何衡量等问题。

基金项目:江苏省自然科学基金项目(BK2004137)
收稿日期:2005-11-16;改回日期:2006-05-30
第一作者简介:赵亚琴(1973~),女。2007 年于南京理工大学获控制科学与工程专业博士学位。主要研究方向为基于内容的信息检索、多媒体技术。E-mail: yaqinzhao@163.com

本文根据视频内容的层次性,提出了建立两种不同粒度信息的层次的电影视频摘要生成方法。视频镜头首先被组织成为镜头类,然后通过分析镜头类之间的时序关系,将表达同一语义的镜头类划分到同一场景。文献[10]利用运动注意力模型检测视频中基于感知的抽象语义信息,为每一个场景选择吸引人的视频片段。本文利用运动注意力模型评价场景中的关键帧和镜头的重要程度,选择场景代表帧和重要镜头,生成的摘要分为镜头级和场景级两个层次。这种两级视频摘要分别由镜头关键帧序列和场景代表帧序列组成,它提供了用户不同粒度的信息,从而方便用户浏览和检索。此外,这样生成的视频摘要由于考虑了人对运动的感知,渗入了人的感知和情感信息,更符合电影或电视视频表达故事情节的方式。

2　视频的场景检测

2.1　关键帧的提取

使用和借鉴文献[11]的镜头边界检测算法,因为它直接在 MPEG 流上进行视频分割,具有很快的检测速度。为了消除镜头渐变的影响,关键帧的提取首先选取镜头中的第 5 帧作为关键帧,顺序计算后继帧与已有关键帧的相似度,如果相似度小于预设的阈值,则该后继帧成为新的关键帧,直到镜头的结束帧。使用颜色直方图[12]来描述帧的视觉内容;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,并且按照色彩的不同范围和主观颜色感知进行不等间隔的量化。定义两帧 f_i, f_j 之间的视觉相似度为

$$FFSim(f_i,f_j)=\sum_{l=1}^{bins}\min(Hf_{il},Hf_{jl}) \quad (1)$$

式(1)中,$bins$ 表示直方图的 bin(盒子)的数目,Hf_{il}, Hf_{jl} 分别是两帧 f_i, f_j 的归一化直方图。设镜头中已有关键帧集合 $KF=\{kf_1,kf_2,\cdots,kf_{N_k}\}$,其中 N_k 为集合中关键帧的数目,定义当前帧 f_i 与已有关键帧集合的相似度为

$$FKSim(f_i,KF)=\max(FFSim(f_i,kf_n)) \quad (kf_n\in KF) \quad (2)$$

2.2　基于滑动镜头窗的视频镜头聚类算法

一个长的视频文档可能包含几千个镜头,这对聚类的速度提出了更高的要求。对于电影视频,一个场景中包含的镜头数目远远少于整个视频,基于此,提出一种基于滑动镜头窗的镜头聚类算法,只需要比较当前窗口内镜头的相似度,不必计算整个视频所有镜头间的相似度,因而,大大减少了镜头相似度的计算次数,提高了聚类速度。此外,由于时序距离很远的镜头不进行相似度比较,这种算法保证了只有同一个场景的相似镜头才被划分到一个镜头类,提高了场景检测的精度。

设 L 指滑动镜头窗的大小(镜头的数目),由于包含在一个场景的镜头数目通常是小于 40,因此 L 的值设定为 40,以保证同一个场景的所有镜头在一个镜头窗内。显然,一个镜头窗中可能包含两个或两个以上场景的镜头。如图 1 所示,镜头窗以 L_w 的增量移动,这里 L_w 是一个可变的量,它的值表征了当前场景的镜头数目。

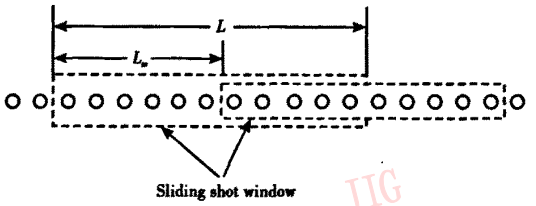

图 1　滑动镜头窗示意图

Fig. 1　Sketch map of sliding shot window

定义 1(滑动镜头窗的移动增量)　设 Sh_i 是滑动窗中的当前镜头,$ShSim(Sh_i,Sh_j)$ 是镜头 Sh_i 和 Sh_j 的视觉相似度,$L_w(i)$ 是 Sh_i 对应的窗口移动增量,则 $L_w(i)$ 用下式计算:

$$L_w(i)=\begin{cases} i & ShSim(Sh_{i+k},Sh_j)>\beta_1 \quad (k=1,2,3;j=1,\cdots,i) \\ 0 & \text{其他} \end{cases} \quad (3)$$

当前镜头窗的移动增量定义为

$$L_w=\begin{cases} L_w(i) & L_w(i)=i \text{ and } L_w(j)=0 \\ L & all\ L_w(i)=0 \end{cases} \quad i=1,2,\cdots,L;\ j=1,2,\cdots,i-1 \quad (4)$$

在电影视频中,一个场景的相似镜头时序距离通常小于 3,如果从某个镜头开始以后连续 3 个镜头与前面的镜头都不相似,那么就认为该镜头是下一个场景的起始。因此,k 的取值为从 1 到 3。

定义两个镜头间的相似度为

$$ShSim(Sh_i,Sh_j)=\max_{kf_k\in KF_i,kf_l\in KF_j}(FFSim(kf_k,kf_l)) \quad (5)$$

其中,KF_i, KF_j 分别是镜头 Sh_i 和 Sh_j 的关键帧集合。定义镜头 Sh_i 和镜头类 SC_k 的相似度为

$$ShSCSim(Sh_i,SC_k)=\max_{Sh_j\in SC_k}(ShSim(Sh_i,Sh_j)) \quad (6)$$

设$|SC_i|$,$|SC_j|$分别表示镜头类 SC_i 和 SC_j 中的镜头数目,且$|SC_i|<|SC_j|$,则定义镜头类 SC_i 和 SC_j 的相似度为

$$SCSim(SC_i,SC_j)=\frac{1}{|SC_i|}\sum_{k=1}^{|SC_i|}ShSCSim(Sh_k,SC_j) \quad (7)$$

基于滑动镜头窗的镜头聚类算法,首先比较当前镜头窗的任意两个镜头(或镜头类)的相似度,如果大于阈值 β_1,则合并为一类。当所有的镜头(或镜头类)比较完后,根据式(3)和式(4)计算窗口的移动增量 L_w,以 L_w 为增量移动镜头窗。

2.3 场景的检测

视频场景是由一组语义相关的镜头组成。根据电影制作和编辑中常用场景特征,视频场景中内容的发展模式可以分为 3 类[12,13]:第 1 类是顺序进展模式,组成场景的各镜头可能发生在同一个地点共享同一背景,在色彩、光照和声音上保持了连贯性,相对于不属于同一场景的镜头,这些镜头之间具有更高的视觉相似度;第 2 类是交错进展模式,组成这种场景的相邻镜头在视觉上可能相似,也可能有很大差别,但他们表达着同一个主题,交替显示;第 3 类是混合进展模式,内容相似的镜头以及交替显示的镜头组合起来表达一个完整的故事情节。顺序进展的场景由一个或几个视觉上相似的镜头类构成;交错进展和混合进展的场景由视觉上相似且时序上交叠的镜头类构成。可见,一个镜头类描述了视频场景的一个侧面或故事线索,而多个镜头类则代表了多个故事线索,这些镜头类在时间维上相互交叠、衔接,共同表达了同一个主题。因此本文通过沿视频流追踪镜头在多个镜头类之间的"跳动"顺序,来判断场景的边界,将表达同一语义的镜头类组织到同一个场景。

定义 2(两个镜头类的时序距离) 设 ID_i^f,ID_i^e 分别是镜头类 SC_i 的最小的镜头序号和最大的镜头序号,定义两个镜头类的时序距离为

$$TD(SC_i,SC_j)=\begin{cases}ID_j^f-ID_i^e & ID_j^f\geqslant ID_i^f \text{ and } ID_j^e-ID_i^f\leqslant L\\ ID_i^f-ID_j^e & ID_j^f<ID_i^f \text{ and } ID_i^e-ID_j^f\leqslant L\\ 0 & ID_j^e-ID_i^f>L \text{ or } ID_i^e-ID_j^f>L\end{cases} \quad (8)$$

如果 $TD(SC_i,SC_j)=0$,那么镜头类 SC_i,SC_j 一定属于不同的场景;如果 $TD(SC_i,SC_j)<0$,是交错关系;如果 $TD(SC_i,SC_j)=1$,两个镜头类是相邻关系;如果 $TD(SC_i,SC_j)>1$,是分离关系。如图 2 所示。

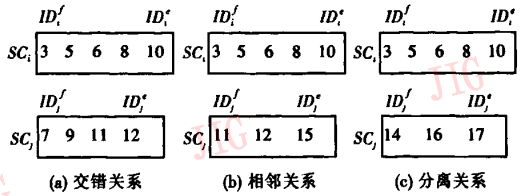

图 2 两个镜头类的时序关系

Fig. 2 Temporal relationship between two shot clusters

场景检测算法的思路是首先合并所有交错关系的镜头类,然后判断任意两个相邻关系的镜头类的相似度是否大于给定的阈值 T_{s_k},如果满足条件,则合并成一个镜头类,否则不合并。最终形成的每个镜头类对应一个场景。

场景检测算法

输入:镜头类序列

输出:场景序列

(1) 初始化,输入镜头类序列

(2) 对于任意的两个镜头类,如果是交错关系,则合并为一个新的镜头类;

(3) 重复第 2 步,直到不再有交错关系的镜头类;

(4) 如果两个相邻关系的镜头类满足条件 $SCSim(SC_i,SC_j)\geqslant T_{s_k}$,则合并;

(5) 重复第 2 步,直到不再有满足条件的相邻关系的镜头类可以合并;

(6) 所有合并产生的新的镜头类和未合并的镜头类都被看作是场景。

3 摘要的生成

场景中的镜头和关键帧在表达故事情节时重要程度不同,摘要中应尽量包含能够反映故事主要内容的重要关键帧,这样更符合人们对摘要需求的心理。运动注意力模型[10]被用来选择场景中的代表帧和重要镜头。

3.1 运动注意力模型

神经生物学上的注意力是指人们仔细观察对象或者聆听的精神力量集中,运动注意力模型通过统计运动向量的密度、空间和时间连续性来模拟视网膜对运动的感知,并组织成运动显著性图检测用户感兴趣或令人兴奋的运动。运动注意力模型把人眼

的视网膜看作一个运动向量场(motion vector field, MVF),假定 MVF 由 3 种注意力感应器组成:强度感应器、空间连续感应器和时间连续感应器。当运动向量通过这 3 个感应器时,分别生成描述强度、空间连续和时间连续的 3 个图像,这些图像被融合成一个运动显著性图来建模人类的注意力。

运动向量的特征直接从 MPEG 域上提取,强度感应器 C_I 帧表征运动的能量或行为;时间、空间感应器 C_s,C_t 分别表征了运动向量的时间、空间连续性。对于每一个宏块,计算 $W\times W$(像素)空间窗内的直方图 $SH_{ij}^{\blacktriangledown}$,以及长度为 L_T 帧的时间窗内的直方图 $TH_{ij}^{L_T}$。则 3 种感应器分别用下面的公式计算[10]:

$$C_I(i,j)=\frac{\sqrt{dx_{ij}^2+dy_{ij}^2}}{MaxMag} \tag{9}$$

其中,$MaxMag$ 指在 MVF 内的最大 C_I 值。

$$C_s(i,j)=-\sum_{k=1}^{bins}P_s(k)\log(P_s(k)) \tag{10}$$

$$C_t(i,j)=-\sum_{k=1}^{bins}P_t(k)\log(P_t(k)) \tag{11}$$

其中,$P_s(k)=\frac{SH_{ij}^{\blacktriangledown}}{\sum_{l=1}^{bins}SH_{ij}^{\blacktriangledown}(l)}$,$P_t(k)=\frac{TH_{ij}^{L_T}}{\sum_{l=1}^{bins}TH_{ij}^{L_T}(l)}$,$bins$ 表示直方图的 bin(盒子)的数目。

这 3 个感应器的输出值以一种特定的方式表征了运动的时空特性。总的来说,高强度的运动总是更吸引人的注意力。I 帧感应器用来检测高强度的运动,但对于低能量的运动是不敏感的;而空间感应器 C_s 通常对于低能量的运动是敏感的;时间感应器 C_t 对于移动的物体是很敏感的,用来从相机运动中识别运动的物体。所以运动注意力模型 MA 定义为[10]

$$MA=I\times C_t\times(1-I\times C_s) \tag{12}$$

通过上面的定义,图像帧中 MA 值高的区域更可能是被关注的行为活动。然后相继应用图像处理方法中的直方图均衡、中值滤波、二值化、区域增长和区域选择方法检测运动注意力区域。考虑到人们的注意力通常不能同时集中在 3 个以上的物体上,选择的运动注意力区域的最大限制是 3。

3.2　摘要的生成

用一个图像帧中所有被选择的运动注意力区域的 MA 值的平均值 $MA(kf_i)$ 表征这个图像帧的被关注程度,$MA(kf_i)$ 称为图像帧的注意力值(attention value),一个镜头 Sh_i 的注意力值 $MA(Sh_i)$ 由它关键帧的注意力值的平均值计算,一个镜头类 SC_i 的注意力值 $MA(SC_i)$ 由它包含的镜头的注意力值的平均值计算。

3.2.1　场景级摘要

场景级摘要是从场景中选择能够反映其主要内容的代表帧集合。考虑到场景通常由多个镜头组成,所以每个镜头只选择一个运动注意力值最大的关键帧。每个场景选择的代表帧数目与场景的重要程度相关。场景的重要程度主要考虑场景中的镜头类的数目和运动注意力值两个因素。假定 r 为预设的总的图像帧比率。一个场景 S_i 的重要程度定义为

$$SQ_i=\frac{1}{n_i}\sum_{SC_j\in S_i}|SC_j|\times MA(SC_j) \tag{13}$$

其中,n_i 为场景 S_i 中镜头类的数目。以 $MA(kf_i)$ 从大到小的顺序排列每一个场景中的所有关键帧,用 $FQ(kf_j)=\frac{MA(kf_i)}{\sum_{kf_i\in S_i}MA(kf_i)}$ 评价场景 S_i 中关键帧的重要度,则累计值满足条件

$$\sum_{kf_j\in S_i}FQ(kf_j)\geqslant(1+r)\frac{SQ_i}{\sum_{S_i}SQ_i} \tag{14}$$

的排在最前面的关键帧被选择,其为该场景的代表帧。这样保证了运动注意力大的关键帧被优先选择为场景代表帧,且一个场景中选择的代表帧数目与场景的重要程度成正比,场景的重要程度越高,在该场景中选择的代表帧数目越多,这符合人们对摘要的要求。

3.2.2　镜头级摘要

由于一个电影视频中包含的镜头数目是相当多的,如果摘要中保留每个镜头的关键帧,这样形成的摘要显得冗长杂乱,并不能很好地反映主题。因此,只选择一些重要的镜头,用这些重要镜头的关键帧集合表达视频。这类摘要生成的思路是首先根据场景中镜头类的重要程度,为每一个场景的镜头类选择相应数目的重要镜头,由这些重要镜头的关键帧集合形成摘要。

对于一个场景 S_i,它的镜头类 SC_j 的重要程度定义为

$$SCQ_j=\frac{|SC_j|\times MA(SC_j)}{\sum_{SC_j\in S_i}|SC_j|\times MA(SC_j)} \tag{15}$$

以 $MA(Sh_i)$ 从大到小的顺序排列每一个镜头类的镜头,用 $ShQ(Sh_i)=\frac{MA(Sh_i)}{\sum_{Sh_i\in SC_j}MA(Sh_i)}$ 评价镜头类 SC_j

中镜头的重要度,则累计值满足条件

$$\sum_{Sh_i \in SC_j} ShQ(Sh_i) \geqslant (1+r)SCQ_j \qquad (16)$$

的排在最前面的镜头被保留,其余的镜头被丢弃。

4 实验结果

为了检验本文中提出的视频场景检测方法和摘要生成方法,选取了以下 4 个具有代表性的视频片段作为实验对象。这 4 个视频片段分别来源于《阿甘正传》、《手足情》、《永无宁日》和《谁与争锋》,如表 1 所示。其中前两个视频片段的视频内容变化较平缓,后两个视频片段都属于动作片,其镜头切换频繁,视频内容变化较剧烈。

表 1 测试视频

Tab. 1 Testing video data

视频片段	镜头	时间	实际场景数
《阿甘正传》	104	19′23″	12
《手足情》	92	24′08″	10
《永无宁日》	175	14′27″	21
《谁与争锋》	203	11′46″	29

4.1 场景的检测结果

基于时序结构图的方法[14]是一种常用的效果较好的场景检测方法,但对于某些不形成回路的同一个场景的镜头类,会误划分到不同的场景中。我们定义了镜头类的相邻关系及划分到同一场景的条件,能够更好地检测顺序进展和混合进展模式的场景,提高场景检测的精度。表 2 给出了文献[14]的方法和本文方法对 4 个视频片段进行场景检测的实验结果。从表 2 中可以看出,本文的场景检测方法的查全率和准确率比文献[14]分别提高了约 5.4% 和 9.3%,但是与视频片段 1 和 2 相比,视频片段 3 和 4 的场景检测准确率相对较低,这是由于这两个片段都属于动作片,一些紧张、激烈的场景内容变化很快,某些不相似的镜头在语义上也可能属于同一个场景,因此,单纯地通过从视觉上分析镜头类的时序关系,不足以获取正确的语义信息。由于这些场景常常伴有声音效果,因此需要融合声音等高层语义特征以提高场景检测精度。

《阿甘正传》的一部分连续视频片段的场景检测结果示例如图 3 所示,该片段共有 13 个镜头,图中两行镜头序列分别表示两个场景,每个场景的镜头序列的上面一行注释文字说明该场景包含的镜头类,以及每个镜头类中的镜头序号。如上面一个场景由 2 个镜头类组成,且第 1 个镜头类包含镜头 1,2 和 4,第 2 个镜头类包含镜头 3,5。

表 2 场景检测算法的实验结果

Tab. 2 Results of two scene detection algorithms

单位:%

视频片段	本文方法		时序结构图方法[14]	
	Recall	Precision	Recall	Precision
《阿甘正传》	91.7	84.6	91.7	80
《手足情》	100	90.9	90	75
《永无宁日》	85.7	83.3	80.9	78.2
《谁与争锋》	89.6	83.9	82.8	72.7
平均	91.75	85.68	86.35	76.35

图 3 《阿甘正传》的部分视频片段的场景检测示例

Fig. 3 Demonstration of scene detection for《Freesand Leo》

4.2 摘要的生成结果

采用文献[10]建议的摘要性能指标 *Informativeness* 评价本文提出的摘要生成方法的内容涵盖。用 *Representative* 评价场景级摘要的代表帧集合和镜头级摘要的关键帧集合是否能够很好地概括表达视频的主要内容。场景级摘要视频总的图像帧比率设定为 25%,镜头级摘要总的图像帧比率设定为 40%。邀请了 10 位同学作为测试对象,以上述两个指标为评分标准,请他们分别为 4 个测试视频片段生成的两种摘要打分,总分为 100 分,表 3 显示了 10 位同学给出的分数的平均分。从表 3 中可以看出,两种摘要都能够较全面地涵盖视频内容,而且所选择的代表帧或关键帧具有较好的代表性。相比较而言,场景级摘要更紧凑精炼,概括性更好;镜头级摘要中包含较多的图像帧,内容涵盖相应地也更全面。但是视频片段 1 和 2 的摘要概括性较低,

这是由于这些片段中的某些感人情节画面内静止的成分较多,只是一些人物的对话等,运动注意力模型是以检测图像帧中的运动注意力区域为基础的,因此,这些重要场景的代表帧被选择的机会较少。

表 3　10 个实验者对摘要的评价

Tab. 3　Performance evaluation of video summaries from 10 students

视频片段	*Informativeness*		*Representative*	
	场景级	镜头级	场景级	镜头级
《阿甘正传》	86.6	88.4	72.1	69.3
《手足情》	83.2	89.7	70.3	67.5
《永无宁日》	80.6	81.4	88.9	78.2
《谁与争锋》	79.1	80.3	86.2	76.2
平均	79.8	84.9	79.4	72.8

5　结　论

根据电影视频的编辑特点,提出了一种基于镜头类时空约束的场景检测算法,在此基础上,利用运动注意力模型选择场景的代表帧和重要镜头,建立了两种层次的电影视频摘要。电影视频的场景分析保证了摘要的较全面涵盖;运动注意力模型用于选择重要的代表帧,使得生成的摘要更紧凑,概括性更好。

参考文献(References)

1　Aya Aner-Wolfa, John R. Kender. Video summaries and cross-referencing through mosaic-based representation[J]. Computer Vision and Image Understanding, 2004, 95(2): 201 ~ 237.

2　Müfit Ferman, A. Murat Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences[J]. IEEE Transactions on multimedia, 2003, 5(2): 244 ~ 256.

3　Hanjalic A, Zhang H J. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis[J]. IEEE Transactions on Circuits System and Video Technology, 1999, 9(8): 1280 ~ 1289.

4　Wang Peng, Ma Yu-fei, Zhang Hong-jiang. An novel video indexing approach based on people-similarity[J]. Acta Electronic Sinica, 2004, 32(6): 968 ~ 972. [王鹏, 马宇飞, 张宏江. 一种利用人物相似度的视频索引算法[J]. 电子学报, 2004, 32(6): 968 ~ 972.]

5　Zhang H J, Zhong D, Smoliar S W. An integrated system for content-based video retrieval and browsing[J]. Pattern Recognition, 1997, 30(4): 643 ~ 658.

6　Wolf W. Key frame selection by motion analysis [A]. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C], Atlanta, GA, USA, 1996: 1228 ~ 1231.

7　Aner A, Kender J R. Video summaries through mosaic-based shot and scene clustering [A]. In: Proceedings of the 7th European Conference on Computer Vision-Part IV table of contents [C], Copenhagen, Denmark, 2002: 388 ~ 402.

8　Uchihashi S, Foote J, Girgensohn A, *et al*. Video mange: Generating semantically meaningly video summaries [A]. In: Proceedings of Association for Computing Machinery, Multimedia'99 [C], Orlando, Florida, USA, 1999: 383 ~ 392.

9　Lu Hai-bin, Zhang Yu-jin, Yang Wei-ping. Shot and episode based nonlinear organization of video[J]. Chinese Journal of Computers, 2000, 23(5): 548 ~ 552. [陆海斌, 章毓晋, 杨卫平. 基于镜头和情节的视频非线性组织[J]. 计算机学报, 2000, 23(5): 548 ~ 552.]

10　Ma Yu-fei, Zhang Hong-jiang. A model of motion attention for video skimming[A]. In: Proceedings of International Conference on Image Process, Rochester[C], New York, USA, 2002, 1: 129 ~ 132.

11　Ngo C W, Pong T C, Chin R T. Video partitioning by temporal slice coherency[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11(8): 941 ~ 953.

12　Cheng Wen-gang, Xu De, Lang Cong-yan. An efficient method for video scene detection[J]. Journal of Image and Graphics, 2004, 9(8): 984 ~ 990. [程文刚, 须德, 郎从妍. 一种有效的视频场景检测方法[J]. 中国图象图形学报, 2004, 9(8): 984 ~ 990.]

13　Zhu X Q, Wu X D, Fan J P, *et al*. Exploring video content structure for hierarchical summarization, Multimedia Systems, 2004.

14　Wang Dong-hui, Zhu Miao-liang, Wu Chun-ming. An analysis method for building sequence structure of video stream[J]. Journal of Image and Graphics, 2000, 5(9): 759 ~ 763. [王东辉, 朱森良, 吴春明. 数字视频流的时序结构图分析方法[J]. 中国图象图形学报, 2000, 5(9): 759 ~ 763.]