

基于用户关注空间与注意力分析的视频精彩摘要与排序

黄庆明^{1),2)} 郑轶佳¹⁾ 蒋树强^{2),3)} 高 文⁴⁾

¹⁾(中国科学院研究生院 北京 100190)

²⁾(中国科学院计算技术研究所 北京 100190)

³⁾(中国科学院智能信息处理重点实验室 北京 100190)

⁴⁾(北京大学信息科学与技术学院 北京 100871)

摘 要 文中提出一种基于用户关注空间与注意力分析的视频内容理解方法,该方法可以有效地获得多通道的视频关注信息,并可使用户根据个性化需求定制视频关注内容,实现视频的高效浏览与访问.首先采用基于二叉层次型结构与分类器选择的音频分类算法将视频中的主要声音类型分类,然后将视频中影响用户注意力的视觉、听觉、时序因素定义为用户关注空间,分别使用相应的中层特征在这三个方面对用户注意力进行表示并计算其关注度,从而在音视频底层特征与高层认知之间建立有机过渡.作者设计了顺序决策融合算法来融合视觉与听觉关注度,生成关注度时序变化曲线并获得精彩摘要.最后使用支持向量回归模型并引入相关反馈机制来实现用户个性化的精彩片段排序.该项工作的特点是通过建立符合人类认知规律的关注度模型并结合相关反馈技术,对视频内容进行类人理解.实验证明,该方法对提取与生成符合用户个性化要求的视频摘要及排序结果具有良好的效果.

关键词 用户关注空间;注意力分析;关注模型;音视频摘要;精彩排序

中图法分类号 TP391

User Attention Analysis Based Video Summarization and Highlight Ranking

HUANG Qing-Ming^{1),2)} ZHENG Yi-Jia¹⁾ JIANG Shu-Qiang^{2),3)} GAO Wen⁴⁾

¹⁾(Graduate University of Chinese Academy of Sciences, Beijing 100190)

²⁾(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

³⁾(Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences, Beijing 100190)

⁴⁾(School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

Abstract This paper proposes a user attention analysis based video content understanding approach, which can be used to automatically detect the highlights of videos and rank them according to their impressive values. Firstly, audio classification is done using the authors' hierarchical bintree framework and classifier selection algorithm. Then, the user attention space is established and the visual, aural, temporal mid-level features are extracted to represent the three main modalities of this space, and the attention values are calculated correspondingly. A specific fusion strategy called ordinal-decision is used to combine the visual, aural attention models and form the attention curve for a video. The highlight segments can be extracted from this attention curve. Finally, the support vector regression model and relevance feedback mechanism are employed to rank the highlight segments and make the ranking result more suitable for human personalization.

收稿日期:2008-07-10. 本课题得到国家“八六三”高技术研究发展计划项目基金(2006AA01Z117)和国家自然科学基金(60773136, 60702035)资助. 黄庆明,男,1965年生,博士,教授,博士生导师,主要研究领域为多媒体技术、图像和视频分析、模式识别、计算机视觉等. 郑轶佳,女,1982年生,硕士,主要研究方向为多媒体内容分析与模式识别等. 蒋树强,男,1977年生,博士,助理研究员,主要研究方向为多媒体处理与语义理解、模式识别、计算机视觉等. 高 文,男,1956年生,博士,教授,博士生导师,主要研究领域为多媒体技术、视频编码、计算机视觉、人工智能等.

The method that introduces the user attention into the video content analysis field could effectively generate the summaries and rank them according to their impressive values. The proposed approach is based on the changes of human attention while watching videos rather than the simple content changes of them, which is more consistent with human understanding. Experimental results demonstrate that the proposed approach is effective for video summarization and highlight ranking.

Keywords user attention space; attention analysis; attention model; audio-video summarization; highlight ranking

1 引言

随着数字视频信息越来越多的涌现,如何自动快速地从海量视频数据中找出符合用户个性化要求的内容已经成为一个亟待解决的问题^[1-4]. 对视频所表达的语义信息进行标注,实现自动的视频内容分析与理解可以减轻人工浏览视频内容的工作量,节省检索时间并降低检索强度,具有很高的研究与应用价值. 另一方面,随着诸如 3G 无线通信等多种新型应用环境的不断涌现,迫切要求视频分析系统包含更强的个性化信息,进而能够根据用户的定制需求进行有针对性的操作,并返回用户最为关心的结果. 对视频内容进行类人思维方式的理,能够使计算机对视频内容分析的结果更贴近用户的要求,从而使计算机视频分析技术更加符合人类认知的思维特点.

对视频按照用户需求进行精彩片段提取与排序,实现个性化的视频浏览、检索以及定制等服务是基于内容的视频理解技术的一个重要研究方向,也具有巨大的学术价值和应用前景,吸引了学术界和产业界等不同领域研究人员的广泛关注^[1-7]. 目前,美国卡内基梅隆大学 Informedia 研究组、哥伦比亚大学数字视频与多媒体实验室、罗彻斯特大学图像处理实验室、荷兰代尔夫特工业大学视频索引与检索实验室、意大利佛罗伦萨大学视觉信息处理实验室、新加坡资讯与通信研究院及国内的微软亚洲研究院、清华大学、中国科学院计算技术研究所等都拥有这方面的研究成果. 一些较为成熟的音视频分析系统也不断被开发出来,如哥伦比亚大学的足球广播视频结构分析系统、罗彻斯特大学的体育视频自动分析软件 ASVA (Automatic Sports Video Analyzer) 及 IBM 的 QBIC/Cuevideo、卡内基梅隆大学的 Informedia 等.

本文的研究是在前人已有工作的基础上,进一步提出解决视频精彩摘要与排序问题的新方法. 根

据著名心理学家 Treisman 提出的人类注意力心理学模型理论,人们对事件的认知会受到来自不同渠道信息源的多种信息影响,各模态信息交互作用,共同决定人们的心理活动情况^[8]. 心理学大师 Posner 等通过研究人脑结构与注意力之间的关系扩展了前人的工作,也提出了将人的注意力系统划分为几个执行不同功能但相互关联的子系统,每个子系统分别代表对人类注意力影响的一个方面^[9].

对于视频,其中的一些片段、场景、物体或音乐、对白等如果比该视频的其它部分内容更加引起观众的兴趣,吸引观众注意力,则可认为该内容具有较高的“关注度”(attention). 视频关注度分析,也称视频注意力分析,就是要从视频中自动将这些用户关注值较高的内容提取出来,以帮助生成摘要及实现精彩片段排序.

在本文中,我们提出了基于用户关注空间与注意力分析的视频内容理解方法,目标是在定义用户关注空间、构建用户关注模型的基础上实现一个可扩展的视频精彩片段摘要与排序系统. 其研究内容涵盖了底层特征提取、中层情感特征表示、高层语义描述等多个方面. 首先我们提出一种基于二叉层次型结构与多分类器选择的音频分类树算法对视频中的主要声音类型进行划分;然后分别提取视觉与听觉中层特征来建立用户注意力模型. 使用顺序决策融合方法融合视觉与听觉注意力模型并获得该视频的注意力时序变化曲线,并从该曲线上定位精彩片段;最后,在提取视频精彩片段的基础上,我们结合心理学理论分析观众对精彩的认知情况,在视觉与听觉关注度的基础上加入时序关注度分析,并将视觉、听觉、时序这 3 个精彩程度评价影响因素构成的集合定义为用户关注空间,通过引入相关反馈技术在该空间中捕捉用户关注区域并获得用户个性化排序结果. 通过建立支持向量回归的精彩排序模型,结合用户的相关反馈来获得符合其个性化偏好的排序结果. 本文的研究框架如图 1 所示.

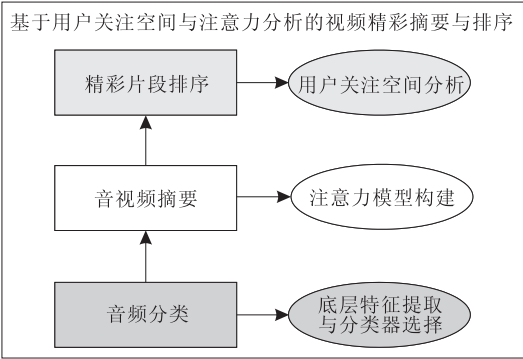


图 1 研究框架图

2 基于二叉层次型结构与多分类器选择的音频分类

每类视频都会有其特定的音频类型. 例如在体育视频中, 观众的欢呼声、解说员的解说声以及比赛相关声音是其主要音频类型. 当前一些通用的音频分类算法一般采用单一层次型结构, 使用固定的分类器与分类特征, 分类效果较差^[10-11]. 为了进行较为准确的音频分类, 为本文后续工作打下良好的基础, 我们提出了一种基于二叉层次型结构与分类器选择的音频分类树算法. 我们以球拍型体育视频中的音频分类问题为例介绍该算法的流程, 如图 2 所示.

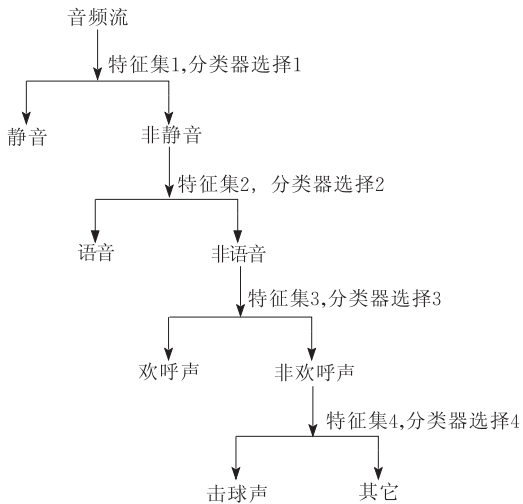


图 2 二叉层次型结构与多分类器选择的音频分类算法

2.1 底层特征选择与提取

以球拍型体育视频为例, 首先对原始音频数据做预加重处理(Re-Emphasize), 以减少尖锐噪声影响, 提升高频信号. 设 $x(n)$ 为原始音频信号, $y(n)$ 为处理后信号, 则

$$y(n) = x(n) - 0.97 \times x(n-1) \tag{1}$$

将处理后的音频分割为定长的音频例子(audio

samples), 相邻音频例子间取 50% 的重叠, 将这些带有重叠的音频例子作为训练与测试的基本单元.

在训练与测试时, 按照分类类别分层进行音频分类. 每一层分别提取不同的底层特征, 其选择依据是选取对该层的两种音频类型经验分类效果最优的底层特征, 并使用不同的分类器进行分类. 在图 2 中, 第 1 层使用短时平均能量与过零率两个特征来区分静音与非静音; 在第 2 层使用 MFCC(12 维)、音调(Pitch)、静音比例、低频能量比率和高过零率比率区分语音与非语音. 由于 MFCC 描述了人耳对频率感知的非线性特征, 所以常被用于语音识别与说话人识别. Pitch 是语音中的音调特征, 是判别语音与非语音的重要特征之一. 静音比例是音频例子特征, 定义如下:

$$silencerate = \frac{silence}{all} \times 100\% \tag{2}$$

即一段音频例子中静音采样点个数占整个音频例子采样点个数的百分比. 由于语音较其它类型声音会有较多的停顿之处, 所以静音比例是区分语音与其它类型声音的良好特征. 低频能量比率为频域音频例子特征, 在非静音音频中, 语音比其它类型声音含有更多静音, 因此语音信号中频域能量低于某个阈值的比例要高于其它类型, 所以该音频例子特征也是区分语音与非语音的一个显著特征. 低频能量比率定义为

$$LERate = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(\text{avg}(E/2) - E(n)) + 1] \tag{3}$$

高过零率比率音频例子特征定义为

$$ZCRRate = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sgn}(ZCR(n) - 1.5\text{avg}ZCR) + 1] \tag{4}$$

以上两式中, N 为一个音频例子中的帧数; $E(n)$ 为第 n 帧的频域能量; ZCR 表示过零率; avg 表示求平均运算, sgn 表示返回括号内函数的整数的运算. 在图 2 中, 第 3 层使用短时平均能量、过零率、带宽来判别欢呼声与非欢呼声. 在欢呼声中过零率的变化率要低于其它类型音频, 所以该特征为区分欢呼声的良好特征之一; 第 4 层使用子带能量、带宽、过零率、频率中心特征来分类击球声与其它类型声音.

2.2 分类器选择

由于采样、样本分布及特征提取方法的不同, 不同的分类器对某个单独特征或某个单独分类问题会表现出不同的分类效果, 所以综合各个分类器的优

点将分类器进行组合与选择来进行决策判别可以提高分类的准确度,达到比使用单一分类器更好的性能。

对二叉层次型音频分类树的每一层,当输入待测试的音频时,在每一层使用的测试分类器均选取在训练时对该层所使用的特征分类效果最好的分类器。对于每一种分类器,其分类效果的通用考量标准是查准率和查全率,分别用下述公式来计算:

$$\text{查全率} = \frac{\text{某音频类型被正确检测到的音频例子数}}{\text{该音频类型总的音频例子数}} \quad (5)$$

$$\text{查准率} = \frac{\text{某音频类型被正确检测到的音频例子数}}{\text{被判定为该音频类型的音频例子数}} \quad (6)$$

为保证查准率与查全率综合最优,我们依据下述原则选择测试分类器:

设使用的分类器集合为 $F = \{F_1, F_2, \dots, F_l\}$, 第 i 层的训练样本集合为 $X_i = \{X_{i1}, X_{i2}, \dots, X_{in}\}$, 第 i 层的音频类别为 $A_i = \{A_{i1}, \bar{A}_{i1}\}$. 对第 i 层上分类器 F_j 的选择方法是

$$F_j = \arg \max \left\{ \frac{P_i(F_j) \times R_i(F_j)}{P_i(F_j) + R_i(F_j)} \right\} \quad (7)$$

其中,

$$P_i(F_j) = \frac{\sum_{k=1}^n (X_{ik} \in A_{i1} \& \& F_j(X_{ik} \in A_{i1}))}{\sum_{k=1}^n (F_j(X_{ik} \in A_{i1}))} \quad (8)$$

$$R_i(F_j) = \frac{\sum_{k=1}^n (X_{ik} \in A_{i1} \& \& F_j(X_{ik} \in A_{i1}))}{\sum_{k=1}^n (X_{ik} \in A_{i1})} \quad (9)$$

在式(7)中 \arg 表示取令括号内分式具有最大值的参数 j 的值; $R_i(F_j)$ 和 $P_i(F_j)$ 分别为按照式(8)、(9)所表示的在第 i 层上分类器 F_j 的查全率与查准率。式(7)的含义是:对该层的两类声音类型,分类效果最优的分类器是使大括号内函数取得最大值的分类器。在测试时使用该分类器对未知数据进行分类可降低运算复杂度,提高运算效率。

本文第5节中,我们在球拍型体育视频上进行音频分类测试来验证该算法的有效性。从实验结果中可以看出,该算法具有良好的分类效果。

3 基于关注度分析的视频摘要

由于视频数据是以非结构化形式存储的,随着

个人数字移动设备中多媒体技术的大量应用,提取其中的精彩片段可以满足人们随时随地浏览视频数据的要求。同时,无线传输设备带宽的限制,要求能够用有限的带宽成本来获得最有价值即最精彩的信息以节省下载花费,视频摘要技术满足了这种移动用户的定制需求。

本节以球拍型体育视频为例,提出一种可扩展的基于关注度分析的视频摘要框架。我们选取符合人类认知规律的中层特征来对音视频关注度进行表示与建模,通过对视频中影响观众注意力的主要因素进行分析来获得关注度变化情况,从而生成摘要。

3.1 听觉关注度分析

从听觉角度出发,我们选择视频中的典型声音类型,采用第2节所述的基于二叉层次型结构与多分类器选择的音频分类树算法对主要声音类型分类,将视频按照时序变化使用不同的音频类型进行标定。

在对各声音类型分类的基础上,本小节对球拍型体育视频中的主要声音类型进行关注度分析。这些主要声音类型中包含了丰富的语义信息,更能吸引观众的注意力。在球拍型体育节目中,观众的欢呼声、解说员的解说声、运动员击球声等为主要声音类型。我们将伴随精彩片段同时发生的声音定义为精彩同步声音,例如网球比赛对打片段中的击球声;另外,有些声音会紧随在精彩片段之后出现,我们称之为精彩异步声音,例如击球得分片段发生之后的解说员激烈解说声和观众的欢呼声等。相应的精彩同步声音模型是对应于精彩同步声音的关注度模型,精彩异步声音模型是对应于精彩异步声音的关注度模型。

一般来说,音频中音量大小的改变、说话人声调高低的变化以及声音的起伏变化情况都会引起收听者的注意。所以音频中影响用户关注度的因素主要有:短时平均能量(*Energy*)、音调(*Pitch*)和平均过零率(*ZCR*)等,其中短时平均能量的大小可以衡量各类声音的强弱程度,音调的高低可以衡量语音的尖锐程度,平均过零率可以衡量音乐的缓急程度。在本节中采用上述主要影响因素对球拍型体育视频中的典型声音类型具有的关注度表示如下:

$$M_{\text{spe}} = \left(\sum_{i=1}^n \text{Energy}_i \right) \left(\sum_{i=1}^n \text{Pitch}_i \right) / n^2 \times 100\% \quad (10)$$

$$M_{\text{che}} = \left(\sum_{i=1}^p \text{Energy}_i \right) / p \times 100\% \quad (11)$$

$$M_{\text{hit}} = \left(\sum_{i=1}^k \text{Energy}_i \right) / k \times 100\% \quad (12)$$

其中, $M_{spe}, M_{che}, M_{hit}$ 是分别对解说员激烈解说声、观众欢呼声和击球声建立的关注度模型. n, p, k 分别是每个音频例子中采样点的数目.

将上述各关注度模型的取值高斯归一化至区间 $[0,1]$ 内. 对于一段确定的视频中的各个声音类型, 连接每段音频例子上的关注度值, 在时序上可获得多条关注度变化曲线: 激烈解说声关注度曲线 C_{spe} , 欢呼声关注度曲线 C_{che} , 击球声关注度曲线 C_{hit} . 这些曲线从不同方面反映了观众对该文件听觉关注度的变化情况.

3.2 视觉关注度分析

视频中的图像特征如颜色、纹理、形状等可以从一帧图像中计算获得, 称为“视频帧内特征”. 与之对应的, 需从连续多帧图像中获得的视频序列特征称为“视频帧间特征”. 由于视频中的精彩片段通常会持续多帧, 单一视频帧的个别情况通常对整段视频关注度情况影响不大, 所以我们从提高运算效率角度出发采用与视频片段精彩程度密切相关的“视频帧间特征”来对视觉关注度建立评价标准.

视觉模态不仅包含空间信息也包含时间信息, 这些信息都会对用户注意力产生影响. 本小节在视觉模态中分别对空间信息与时间信息进行关注度表示. 在之前的工作中, 平均运动向量(motion vector)矢量值特征可以较好地表征视频帧间的运动情况, 当一秒钟内的平均运动向量矢量值较大时该视频场景往往具有较大的运动变化强度, 更易吸引观众的注意力. 尽管有时运动向量并没有真实地反映视频中的运动信息, 但利用这一特征可以在绝大多数情

况下降低运算复杂度并得到正确结果. 在本小节中把视觉空间关注度 M_{spa} 表示为

$$M_{spa} = \left(\sum_{i=1}^k |MV_i| \right) / k \times 100\%$$

(13)

其中 $|MV_i|$ 表示从解码过程中获得的第 i 帧的平均运动向量的矢量值, k 为视频帧率(例如: 25 帧/s).

对时间维度, 镜头转换率(shot change rate)通常被用于描述摄像机运动. 在球拍型体育节目中, 当镜头切换较为频繁时, 通常是比赛内容紧张激烈的时刻, 观众的注意力也更容易被吸引. 视觉时间关注度 M_{tem} 表示为

$$M_{tem} = e^{((1-(n(k)-p(k))/\delta))} \times 100\%$$

(14)

其中 $p(k)$ 和 $n(k)$ 分别是第 k 帧左右两侧最近邻的镜头边界帧号; 参数 δ 为常数, 由 $n(k)-p(k)$ 确定, 用于保证使 M_{tem} 的值分布于 $0 \sim 100\%$ 之间.

类似地, 各视觉精彩程度评价公式的取值范围也使用高斯归一化标准限定在区间 $[0,1]$ 内, 对于一段确定的视频, 可以使用式(13)、(14)在时序上分别获得两条视觉关注度曲线: 视觉空间关注度曲线 C_{sc} 和视觉时间关注度曲线 C_{tc} .

3.3 顺序决策融合算法与视频摘要生成

在上一节中得到的各视觉与听觉关注度评价公式的取值范围均被高斯归一化标准限定在区间 $[0,1]$ 内, 对于一段确定的视频, 可以使用式(10)~(14)在时序上获得若干条视觉与听觉关注度变化曲线: 视觉空间关注度曲线 C_{sc} , 视觉时间关注度曲线 C_{tc} , 解说声关注度曲线 C_{spe} , 欢呼声关注度曲线 C_{che} 以及击球声关注度曲线 C_{hit} . 这些曲线从不同方面反映

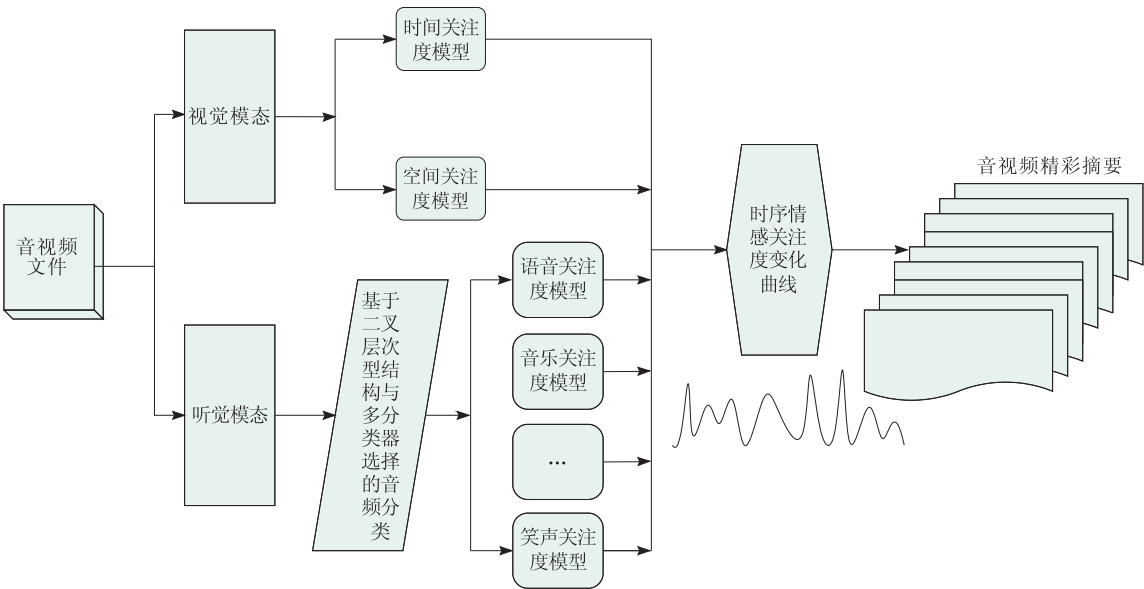


图 3 可扩展的基于关注度分析的视频精彩摘要算法

了观众观看该视频时的关注度变化情况. 在本节中将使用顺序决策融合算法将多条曲线进行融合得到最终的球拍型体育视频关注度时序变化曲线, 并设计相应的算法从该曲线上获得此视频的精彩摘要片段, 其流程如图 3 所示.

3.3.1 顺序决策融合算法

为将多类视觉与听觉关注因素融合成一个整体的评价标准, 我们设计了顺序决策融合算法来实现这一功能. 以球拍型体育视频为例, 顺序决策融合算法表述如下:

首先使用精彩异步曲线 C_{spe} 与 C_{che} 粗略定位发生在解说员激烈解说声与观众欢呼声之前的精彩片段位置, 再利用语音边界检测(静音检测)来精确定位这些精彩片段的左右边界. 使用精彩同步模型 M_{spa} 、 M_{tem} 与 M_{hit} 结合精彩异步模型 M_{spe} 和 M_{che} 来确定最终的球拍型体育视频时序精彩程度变化情况. 采用该顺序决策融合算法所获得的精彩程度评价标准如下式所示:

$$M_a = (\lambda_{spa} \cdot M_{spa} + \lambda_{tem} \cdot M_{tem} + \lambda_{hit} \cdot M_{hit}) \cdot e^{(\sum_{i=1}^p M_{spe})} \cdot e^{(\sum_{i=1}^q M_{che})} \times G(n) \tag{15}$$

其中 λ_{spa} 、 λ_{tem} 、 λ_{hit} 分别为每个精彩同步模型的权重, 满足均大于 0 且 $\lambda_{spa} + \lambda_{tem} + \lambda_{hit} = 1$ 的约束条件. p 、 q 分别为精彩异步模型, 即激烈解说声模型和欢呼声模型的持续时间(以秒为单位). $G(n)$ 为高斯平滑窗, n 是平滑参数.

该算法可用图 4 表示.

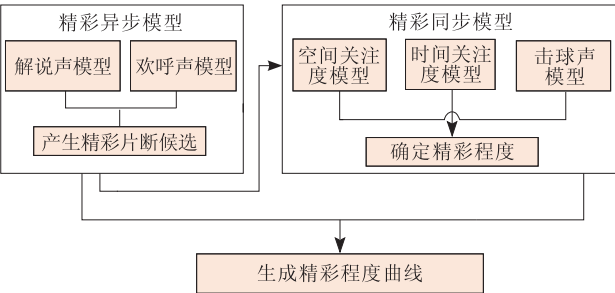


图 4 顺序决策融合算法

3.3.2 视频摘要生成

基于关注度分析的视频摘要算法可表述为: 首先对球拍型体育视频进行音频关注度分析, 然后使用解说声和欢呼声这两种精彩异步因素粗略定位其之前的精彩片段位置. 具体方法是: 以精彩异步因素发生之前的边界为精彩候选片段的右边界, 从该处开始向前查看. 将其之前的击球声片段的开始点设为精彩候选片段的左边界. 最后利用镜头边界检测

结合语音边界检测(静音检测)来精确定位这些精彩片段的左右边界^[12].

本文第 5 节给出了使用基于关注度分析算法在球拍型体育视频中提取精彩摘要的实验结果. 从实验结果中可以看出, 使用该算法提取的精彩摘要较为符合人的主观认知规律.

4 基于用户关注空间与相关反馈技术的视频精彩片段排序

提取视频中的精彩事件并按照精彩程度对其进行排序可以有效地节省观众的浏览时间与下载花费. 因此, 精彩排序技术在基于内容的视频分析领域具有重要的研究价值.

尽管目前已存在一些视频精彩排序方法, 但如何构建一个符合用户个性化要求, 具有在线学习能力的系统已成为一个亟待解决的问题. 为此, 我们提出了一种基于用户关注空间相关反馈技术的可扩展的视频精彩排序方案. 通过分析精彩事件在音频、视频、时间 3 个用户关注子空间上的情感特征, 按照其激烈程度和用户的相关反馈结果来对其进行排序, 以使排序结果更为符合人类感知特性, 满足用户个性化需求.

4.1 用户关注子空间划分与表示

根据心理学家 Treisman 的理论, 观众对视频中精彩事件的理解会受到来自不同渠道的多方面因素的影响. 视觉、听觉、时间信息都是视频中的基本模态, 每一个模态都会对观众关于精彩事件的感知产生重要的影响^[8]. 在我们的精彩排序系统中, 将这种多模态感知方式定义为用户关注空间, 它由 3 个子空间构成: 视觉子空间、听觉子空间和时序子空间, 每一个子空间分别表示观众对该视频中精彩事件不同模态的理解情况.

在该关注空间中用户有其各自对 3 个子空间的偏爱比例, 即不同的用户会按照自身在音频、视频、事件持续时间上的偏好程度来对同一音视频事件给出自己的精彩程度评价. 例如有些观众认为持续时间最长的事件片段是最精彩的, 而有些观众会把具有最强观众欢呼声的事件片段作为最精彩事件.

为了对上述 3 个用户关注子空间进行有效的建模, 我们以球拍型体育视频为例提取有效的情感特征来表示每一个用户关注子空间. 运动情况是一种能够反映事件激烈程度的被广泛应用于体育视频分析领域中的有效的视觉特征, 摄像机运动及运动员

活动如摄像场景变换和运动员轨迹均可在实际分析中加以应用. 在本文中我们提取 MPEG-7 运动活动性描述子^[13]和运动员运动方向转换率^[14],从摄像机运动与运动员运动轨迹的角度对视觉关注子空间进行表示. 运动员的方向转换率越高、MPEG-7 运动活动性描述子越大,则表示某一结构性事件的精彩程度越高. 特别地,在球拍型体育视频中运动方向转换率可以使用对运动员运动分析得出的左右挥拍转换率^[14]来代替. 音频能量和音调相关特征是在体育视频分析中被广泛应用的音频情感特征^[13,15-16],通常欢呼声的平均能量和解说声的平均音调越高说明事件的精彩度越高. 精彩事件的持续时间是另一个可以反映事件精彩程度的有效的情感特征. 以网球运动为例,比赛场景的持续时间越长一般说明运动员的相持越激烈,比赛越精彩. 这一特征在足球等场地运动中可被替换为镜头切换长度特征^[17].

在用户关注空间分析算法中我们共提取 7 个中层特征来对用户关注子空间进行表示,它们是: MPEG-7 平均运动向量、运动方向转换率、欢呼声平均能量、激烈解说声平均音调、精彩事件持续时间、欢呼声持续时间和激烈解说声持续时间. 这 7 个特征已在前人工作中被证明在体育视频分析研究中是非常有效的^[13,15-17]. 使用上述特征,用户关注子空间的划分如表 1 所示.

表 1 用户关注子空间划分与表示

视觉子空间: {MPEG-7 平均运动向量、运动方向转换率}
听觉子空间: {欢呼声平均能量、激烈解说声平均音调}
时间子空间: {精彩事件持续时间、欢呼声持续时间、激烈解说声持续时间}

4.2 用户关注模型建立

按照预定义的 3 个用户关注子空间,可使用支持向量回归方法对视频建立 3 个用户关注子模型(user attention submodel)并产生最终的精彩排序模型. 这 3 个子模型是:视觉子模型、听觉子模型、时序子模型. 它们分别体现了不同观众在观看同一视频时对 3 个用户关注子空间的不同偏爱程度. 我们将最终的个性化精彩排序模型定义为对 3 个子模型的线性求和,在下文中称为联合模型.

我们对视觉、听觉、时序 3 个子空间分别用支持向量回归方法来建立子模型. 支持向量回归(Support Vector Regression, SVR)是一种非线性回归技术,它具有需要较少训练样本、有良好泛化能力的优点,尤其对稀疏非线性的数据分布更具有良好的鲁棒性和预测准确能力^[18]. 所以我们使用 SVR 技术进行建模.

SVR 模型的输入是在 3 个用户关注子空间上提取的中层特征,输出是计算机自动求得的特定观众对各精彩片段的个性化精彩程度评价.

我们将精彩排序模型定义为 3 个关注子空间上各子模型的线性求和形式,可用下式表示为

$$M_{\text{rally}}(s) = \omega_{\text{VSM}} \cdot M_{\text{VSM}}(s) + \omega_{\text{ASM}} \cdot M_{\text{ASM}}(s) + \omega_{\text{TSM}} \cdot M_{\text{TSM}}(s) \tag{16}$$

其中 $M_{\text{VSM}}(s)$, $M_{\text{ASM}}(s)$, $M_{\text{TSM}}(s)$ 分别表示在 3 个用户关注子空间上建立的精彩排序子模型; ω_{VSM} , ω_{ASM} , ω_{TSM} 分别表示 3 个子模型线性求和的权值; $M_{\text{rally}}(s)$ 是由计算机结合用户的相关反馈结果求得的精彩事件的精彩程度值.

4.3 相关反馈捕捉用户关注区域

为在 3 个用户关注子空间中有效捕捉到用户的感兴趣区域,我们引入了相关反馈技术. 相关反馈技术能够使(检索、排序)系统逐步理解用户的个性化查询、检索、排序等请求,并将最符合用户需要的结果反馈给用户.

基于相关反馈的排序过程详述如下:

1. 将 3 个用户关注子空间的初始权值设为 $\{1/3, 1/3, 1/3\}$,使用式(16)计算各精彩片段的初始精彩程度值,将各精彩度值降序排列.
2. 选择前 M 个精彩度值大于 $conf$ 的精彩片段作为初始结果集 $HI = \{s_1, s_2, \dots, s_M\}$, $1 \leq i \leq M$,反馈给用户. s_i 表示被选中的一个精彩片段反馈结果, M 和 $conf$ 由用户输入产生. 观众从 HI 中选择其满意的精彩片段产生观众满意结果集 $HS = \{s'_1, s'_2, \dots, s'_N\}$, $s'_i \in HI$, N 表示在反馈结果中观众满意的精彩片段个数. 排序准确率定义为 $RA = N/M \times 100\%$.
3. 使用视觉子模型结合提取的视觉情感特征来重新分别计算每个 $s_i \in HI$ 的值,获得精彩片段集合 $VSH = \{s_1^V, s_2^V, \dots, s_p^V\}$, 其中 $s_i^V \in HS$ 并且 $M_{\text{VSM}}(s_i^V) \geq conf$.
4. 分别使用听觉子模型结合音频情感特征和时间子模型结合时间情感特征重复步 3, 分别获得精彩片段集合 $ASH = \{s_1^A, s_2^A, \dots, s_q^A\}$ 与 $TSH = \{s_1^T, s_2^T, \dots, s_r^T\}$.
5. 按照 $\{P/(P+Q+R), Q/(P+Q+R), R/(P+Q+R)\}$ 返回步 1 重新设置 3 个子模型的权重,重复步 2 计算新一轮的排序准确率 RA' .
6. 如果 $|RA' - RA| \leq threshold$ 或已得到用户满意的结果,则停止;否则返回步 3.

使用上述算法,每位观众都能够获得依照自身个性化需求求得的权重集 $\{\omega_{\text{VSM}}, \omega_{\text{ASM}}, \omega_{\text{TSM}}\}$,这一权重集体现了该观众在关注空间中对精彩事件的自身感知与理解. 我们的系统可以为每名用户提供一套与之对应的权重集,使用这一权重集可以提供给该用户最符合其自身个性化需求的排序结果. 从算法中可以看出,我们的精彩排序算法具有良好的在

线学习能力,相比于以往传统的视频精彩排序技术来说具有很大的优势^[19].

在本文第5节我们给出了使用基于用户关注空间与相关反馈技术算法对球拍型体育视频进行精彩排序的实验结果.

5 实验结果

为验证本文3个典型算法,我们以球拍型体育视频为研究对象设计了3个相关实验.本实验从2004年奥运会乒乓球比赛与2005年法网公开赛中选择3段网球视频(视频段1~3)与3段乒乓球视频(视频段4~6)作为实验数据,详见表2.

表 2 实验数据	
视频段	时间/s
1	505
2	739
3	574
4	607
5	580
6	676

5.1 基于二叉层次型结构与多分类器选择的音频分类实验结果

如表3所示,我们选择网球视频段1与乒乓球视频段4作为训练数据.其它视频段作为测试数据.

表 3 训练数据						
视频段	时间/s	数目				
		类别 1	类别 2	类别 3	类别 4	类别 5
1	505	114	127	101	98	65
4	607	146	159	125	103	74

注:类别1表示击球声;类别2表示欢呼声;类别3表示静音;类别4表示解说声;类别5表示其它声音类型.

在网球、乒乓球比赛等体育视频中,实验效果与测试音频片段的纯净度、底层特征的选择、分类的类别数目、分类树上各类别层次的先后分类顺序、每层分类器的选择均有关系.表4和表5分别是在网球和乒乓球视频上测试的结果,从中可以看出,采用基于二叉层次型结构与分类器选择的音频分类树算法在查准率与查全率方面均能得到良好的分类效果.

表 4 网球视频测试结果								
类别	时间/s	分类个数					查全率/ %	查准率/ %
		1	2	3	4	5		
1	312	247	31	4	19	11	79.2	83.2
2	359	23	308	6	11	11	85.8	87.0
3	285	8	2	266	3	6	93.3	95.3
4	266	12	8	2	231	6	86.8	88.5
5	91	7	5	1	5	73	80.2	68.2

表 5 乒乓球视频测试结果								
类别	时间/s	分类个数					查全率/ %	查准率/ %
		1	2	3	4	5		
1	268	215	18	6	16	13	80.2	83.0
2	309	12	274	3	11	9	88.7	85.4
3	254	6	4	239	2	3	94.1	93.0
4	223	5	14	4	192	8	86.1	83.8
5	202	21	10	5	8	158	78.2	82.7

5.2 基于关注度分析的视频摘要实验结果

为验证该算法的有效性,我们选用上述音频分类实验中使用的6段球拍型体育视频作为研究对象.此类视频具有受众广泛、形式多样且人工编辑成份较少,所需前期处理工作简单的特点.

在本实验中我们使用摘要精度*accuracy*作为评测标准,摘要精度的计算方法是,令4位具有丰富球拍型体育视频比赛知识的观众独立观看每段视频,之后给出其对使用我们的算法自动提取的视频摘要片段的满意度(分为从1~10个等级),4人对每段视频摘要满意度的均值即为该段摘要的摘要精度.实验结果如表6所示.

表 6 视频摘要算法实验结果		
视频	摘要个数	<i>accuracy</i>
1	21	9.0
2	26	8.75
3	23	9.5
4	28	9.75
5	26	8.5
6	34	9.25

实验证明该方法能有效提取视频中与观众主观认知规律较为一致的精彩摘要片段,不失为一种效果良好的视频摘要生成算法.

5.3 基于用户关注空间与相关反馈技术的视频精彩片段排序实验结果

我们设计了一个基于相关反馈交互的精彩排序原型系统以验证我们的算法.使用网球视频段1、2、3作为测试数据,首先按照第3节的精彩摘要算法检测精彩事件片段;由于精彩是一个主观性很强的概念,我们随后安排了一个主观测试实验以确保实验数据集的随机性.我们邀请4位具有丰富球拍运动知识的观众对每个精彩片段按照他们的个人理解独立进行手工精彩度标注,每一片段均被赋予0.1~1.0之间的一个精彩程度值.比赛片段的精彩度越高其得分也就越高,最终每一片段的得分被定义为4位标注者各自评分的均值.这样,就获得了每一精彩片段的主观评价基准.在实验中我们在上述的6段视频中任意选择其中一段作为支持向量回归模

型的训练样本,其余 5 段作为测试集,共进行两次对比实验来对我们提出的方法进行验证.

为了与以前的相关工作进行比较,首先将提取的所有情感特征作为输入,使用单一模型进行排序.将 4.3 节中算法的参数设为 $M=20, conf=0.7$, 获得前 20 个精彩度值大于 0.7 的最精彩的比赛片段.邀请 4 位观众分别进行精彩打分,通过 4 次计算获得的平均 RA 为 $RA1=81.7\%$ (RA 为算法中定义的排序准确率,表示系统返回的精彩片段中被用户接受的百分比).计算这 $M=20$ 个由用户主观评价打分获得的基准与计算机算得的结果之间的差异,定义为人机差异 dif :

$$dif = \sqrt{\left[\sum_{i=1}^M (s_{ai} - s_{gi})^2 \right] / M} \quad (17)$$

s_{ai} 表示计算机对第 i 个精彩片段的打分, s_{gi} 表示相应的用户主观评价基准.单一排序模型的 4 次人机差异的均值为 $dif1=15.5\%$. dif 有效地反映了计算机与人们主观感知之间的差异性.

为了与上述实验结果比较,我们使用联合模型进行对比实验.设置 3 个子空间的初始权值为 $\{1/3, 1/3, 1/3\}$, 与上次实验一样,输出前 20 个精彩程度得分高于 0.7 的片段并让用户找出在这 20 个片段中能够令他们满意的片段并统计其个数.这样就获得了使用联合模型且不使用相关反馈的实验结果: $RA2=74.3\%$, $dif2=16.9\%$.

在第 2 个对比实验中我们在联合模型中引入了相关反馈.每名观众只需要在系统输出的结果中找出他们满意的精彩片段反馈给计算机系统,系统通过用户的反馈调整子模型的权值,并将使用新权值求得的新一轮精彩程度值中前 20 个比赛片段和它们相应的精彩度打分返回给用户.在本次实验中经 5 次反馈之后系统达到了平均 $RA3=96.4\%$, $dif3=11.5\%$ 的结果.

从实验中可以看出,不引入相关反馈的联合模型的排序效果并不尽如人意,其排序精确率比单一模型要低而平均人机差异值要高于单一模型 ($RA1 > RA2$, $dif1 < dif2$);但是经过若干次相关反馈之后的联合模型在达到稳定后的基于相关反馈的联合模型的排序效果明显优于单一模型 ($RA1 < RA3$, $dif1 > dif3$).对实验中的 4 个特定用户,在 5 次相关反馈交互之后最终的平均 RA 达到 96.4% .本实验证明了相关反馈技术能有效地在用户关注子空间中捕捉到用户的感兴趣区域.

6 结 语

本文研究基于用户关注空间与注意力分析的视频内容理解技术,给出了用户关注空间的定义,并依据视频的特点将用户关注空间划分为多个子空间,在每个子空间上分别提取中层情感特征建立相应的关注度模型,融合各用户关注子模型形成最终的用户关注模型,分析该模型可获得整个视频的时序情感变化情况.

对于本文中基于用户关注空间与注意力分析的视频内容理解技术所涉及的音频分类、视频摘要、视频精彩片段排序方面的应用,本文均给出了实验结果.实验证明,相对于传统的视频理解技术,我们提出的从主观认知角度出发,对视频中蕴含的语义信息进行分析的方法是可行且有效的,这种方法结合心理学中的相关理论,将认知因素考虑到视频分析问题中来,能够更好地解决与用户主观感受有关的视频理解问题.

参 考 文 献

- [1] Hanjalic A, Xu L. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 2005, 7(1): 143-154
- [2] Kang Hang-Bong. Affective content detection using HMM// *Proceedings of the ACM International Conference on Multimedia*. Berkeley, CA, USA, 2003: 259-262
- [3] Kang Hang-Bong. Affective contents retrieval from video with relevance feedback// *Proceedings of the ACM International Conference on Multimedia*. Berkeley, CA, USA, 2003, 2911: 243-252
- [4] Xu Min, Chia Liang Tien, Jin Jesse. Affective content analysis in comedy and horror videos by audio emotional event detection// *Proceedings of the IEEE International Conference on Multimedia and Expo*. Amsterdam, Holland, 2005: 622-625
- [5] Ma Y F, Hua X S, Lu L, Zhang H J. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia*, 2005, 7(5): 907-919
- [6] You J, Liu G, Sun L, Li H. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Transaction on Circuits and Systems for Video Technology*, 2007, 17(3): 273-285
- [7] Wang H, Cheong L. Affective understanding in film. *IEEE Transactions on Circuits and Systems for Video Technology*, 2006, 16(6): 689-704
- [8] Treisman M, Gelade G. A feature-integration theory of attention. *Cognitive Psychology*, 1980, 12: 97-136

- [9] Posner M I, Petersen S E. The attention system of the human brain. *Annual Review of Neuroscience*, 1990, 13: 25-42
- [10] Cai Rui, Lu Lie, Hanjalic Alan, Zhang Hong-Jiang, Cai Lian-Hong. A flexible framework for key audio effects detection and auditory context inference. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, 14 (3): 1026-1039
- [11] Lu Lie, Hanjalic Alan. Towards optimal audio "Keywords" detection for audio content analysis and discovery//*Proceedings of the 14th ACM International Conference on Multimedia*. Santa Barbara, USA, 2006: 825-834
- [12] Zheng Yi-Jia, Zhu Guang-Yu, Jiang Shu-Qiang, Huang Qing-Ming. Visual-aural attention modeling for talk show video highlight detection//*Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Ohio, USA, 2008: 2213-2216
- [13] Xing L, Yu H, Huang Q, Ye Q, Divakaran A. Subjective evaluation criterion for selecting affective feature and modeling highlight//*Proceedings of SPIE Conference on Multimedia Content Analysis, Management*. San Jose, CA, USA, 2006, SPIE-6073: 188-195
- [14] Zhu G, Xu C, Huang Q, Gao W, Xing L. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game//*Proceedings of the ACM International Conference on Multimedia*. Santa Barbara, USA, 2006: 431-440
- [15] Hanjalic A. Generic approach to highlights extraction from a sports video//*Proceedings of the IEEE International Conference on Image and Processing*. Barcelona, Spain, 2003: 1-4
- [16] Xiong Z, Radhakrishnan R, Divakaran A. Generation of sports highlights using motion activity in combination with a common audio feature extraction framework//*Proceedings of the International Conference on Image Processing*. 2003, 1: 5-8
- [17] Tong X, Lu Q, Zhang Y, Lu H. Highlight ranking for sports video browsing//*Proceedings of the ACM International Conference on Multimedia*. Singapore, 2005: 519-522
- [18] Cristianini N, Shawe Taylor J. *An Introduction to Support Vector Machines*. Cambridge, UK: Cambridge University Press, 2000
- [19] Zheng Yi-Jia, Zhu Guang-Yu, Jiang Shu-Qiang, Huang Qing-Ming. Highlight ranking for racquet sports video in user attention subspaces based on relevance feedback//*Proceedings of the IEEE International Conference on Multimedia and Expo*. Beijing, China, 2007: 104-107



HUANG Qing-Ming, born in 1965, Ph. D., professor and Ph. D. supervisor. His research interests include multimedia content analysis, image and video processing, computer vision and pattern recognition.

ZHENG Yi-Jia, born in 1982, MS. Her research interests include multimedia content analysis, pattern recognition.

tion.

JIANG Shu-Qiang, born in 1977, Ph. D., assistant researcher. His research interests include multimedia processing and semantic understanding, pattern recognition, computer vision.

GAO Wen, born in 1956, Ph. D., professor, Ph. D. supervisor. His research interests include multimedia technologies, video coding, computer vision, artificial intelligence.

Background

With the emergence of more and more digital video information, fast and automatic extraction of user oriented personalized video summarization from massive video database has become an issue to be solved. On one hand, by labeling the semantic information expressed by the video and realizing the automatic video content analysis and understanding, we can reduce the work load of manual browsing for video content and save retrieval time, which is of great value in research and application. On the other hand, with the unceasing emergence of many kinds of new application scenarios such as 3G wireless communication environment, a good video analysis system calls for stronger personalized information, so that we can carry on the target-oriented operation ac-

cording to the user's personal demand, and return the results most concerned by the user. If we make computer to understand the video content in a human-being way, we can draw closer to the user's request, thus will cause the video parsing technique to conform better to the characteristics of human perception. In this regards, based on the previous work, the authors further propose the new methods to solve video summarization and highlight ranking issues, trying to understand the video content based on user attention space and attention analysis. The target is to realize an expandable video summarization and highlight ranking system based on the definition of user attention space and construction of user attention model.