计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(8):14−21 [doi: 10.15888/j.cnki.csa.008003]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于语义的视频检索技术综述①

黄　立, 朱定局

(华南师范大学 计算机学院, 广州 510631)
通讯作者: 朱定局, E-mail: zhudingju@m.scnu.edu.cn

摘　要: 本文综述了基于语义的视频检索的研究现状, 以帮助未来的研究人员了解基于语义的视频检索领域中可用的技术, 视频检索系统的产生是为了在互联网或数据库中的大量视频数据集中找到用户想要查询的视频. 本文对基于语义的视频检索过程进行了说明与讨论, 本文还对基于语义的视频检索中, 解决语义鸿沟这一主要问题的相关技术进行了综述. 语义鸿沟的形成是因为从视频内容中提取的低层特征与现实世界中用户对这些特征的认知存在差异, 将视频内容的低层特征转化为高层的语义概念是一个备受关注的研究课题.

关键词: 特征提取; 语义鸿沟; 用户查询; 视频挖掘; 视频检索

引用格式: 黄立,朱定局.基于语义的视频检索技术综述.计算机系统应用,2021,30(8):14–21. http://www.c-s-a.org.cn/1003-3254/8003.html

Review on Semantic-Based Video Retrieval Technology

HUANG Li, ZHU Ding-Ju

(School of Computer Science, South China Normal University, Guangzhou 510631, China)

Abstract: This study summarizes the current research on semantic-based video retrieval to help future researchers understand the technologies available in this field, and video retrieval systems are created to find the video that users want to query in a large number of video data collections on the Internet or in databases. This study introduces and discusses the semantic-based video retrieval process and also summarizes the relevant techniques to solve the main problem of a semantic gap in this process. The semantic gap is induced by the difference between the low-level features extracted from video content and the user's cognition of these features in the real world. It is a highly concerned research topic to transform the low-level features of video content into high-level semantic concepts.

Key words: feature extraction; semantic gap; user query; video mining; video retrieval

随着 5G 网络技术的发展和视频拍摄以及创作技术门槛的降低, 包括以哔哩哔哩为代表的长视频平台和以抖音为代表的短视频平台的视频规模、投稿数和用户活跃度都得到了极速的增长, 导致了现在互联网上的视频数据量呈爆炸式增长. 以长视频平台哔哩哔哩为例, 根据哔哩哔哩 2020 年第二季度的财报显示, 该平台视频创作者月均投稿量相比上个季度同比增长 148%, 日均视频播放量达到了 12 亿次. 面对大量的视频数据, 如何从这些视频库中检索出人们所需的视频, 是当下面临的一个挑战. 因此, 许多视频检索系统也由此而诞生和引入.

本文旨在综述基于语义的视频检索方法, 在第

① 基金项目: 中国高等教育学会专项课题 (2020JXD01); 广东省普通高校“人工智能”重点领域专项 (2019KZDZX1027); 广东高校省级重点平台和重大科研项目 (2017KTSCX048); 广东省中医药局科研项目 (20191411)
Foundation item: Special Project of China Association of Higher Education (2020JXD01); Special Project of Artificial Intelligence for Ordinary Universities of Guangdong Province (2019KZDZX1027); Provincial Major Science and Technology Research Program and Key Platform of Higher Education of Guangdong Province (2017KTSCX048); Research Project of Traditional Chinese Medicine Bureau of Guangdong Province (20191411)
收稿时间: 2020-11-10; 修改时间: 2020-12-12; 采用时间: 2020-12-18; csa 在线出版时间: 2021-07-31

1 节中诠释了相关视频术语, 在第 2 节中讨论了基于语义的视频检索系统的结构, 在第 3 节中对基于语义的视频检索领域中的应用进行了概述, 在最后第 4 节中作了总结与展望.

1 视频检索技术相关概念介绍

视频检索技术的相关概念包括视频检索技术本身的分类和发展, 以及视频的基础概念知识.

1.1 视频检索技术概念

视频检索的检索技术主要有两种形式: 基于文本的视频检索技术 (Text Based Video Retrieval, TBVR)[1] 和基于内容的视频检索技术 (Content Based Video Retrieval, CBVR)[2]. 在基于文本的视频检索技术中, 需要对视频进行大量的手工注释, 这种方法的视频检索依赖于与每个视频相关的元数据, 例如标签、标题、描述和关键字等, 缺点是需要人工进行注释. 基于内容的视频检索技术的研发初衷就是为了解决基于文本的视频检索技术中的缺点, 基于内容的视频检索技术能够自动地识别视频中内容的特征, 例如颜色、纹理、形状等, 然后根据所提取的特征做进一步的处理, 包括关键帧检测提取、聚类和建立索引等工作.

语义表达是构建高效视频数据索引的基础, 除了视频画面中所表现的各种物体颜色和形状等信息, 真正能够让人们识别视频的关键因素还是视频所表达的意义和概念. 因此, 基于语义的视频检索技术 (Semantic Based Video Retrieval, SBVR)[3,4] 是视频检索系统领域的重要研究方向. 通常情况下, 人类能够准确感知视频中的内容所表达的意义, 但计算机的感知能力还远不如人类般切实, 这种差异化的表现被称为语义鸿沟 (semantic gap)[5,6]. 基于语义表达技术的核心思想是将从视频的内容中提取到的低层特征与人类对这些特征的认知理解之间进行映射匹配, 结构如图 1 所示.

1.2 视频概念

视频的属性信息可以分为 3 类: 第 1 类是颜色、形状等视觉上可见的低层特征信息; 第 2 类是听觉上的如响度和音调等, 或是文字和符号等描述信息; 第 3 类是用户能够感知到的视频中发生的事情的语义信息. 能被用来确定视频中所发生的事件的语义的信息包括: 事件对象信息、空间信息和时间信息. 提取不同模态的视频特征的目的, 就是为了弥合低水平特征和高水平语义概念之间的鸿沟.

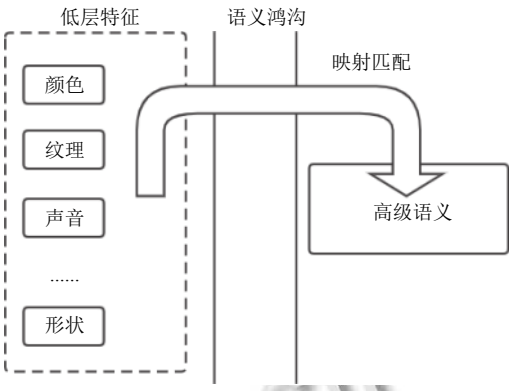

图 1 跨越语义鸿沟

视频的结构自顶向下主要分为: 视频、场景、镜头和帧, 如图 2 所示. 视频是由许多场景组成, 是一组连续静态图像的序列, 同时叙述一个完整的故事结构. 场景是一组在语义上相关、在时间上相邻的镜头, 是在相同的地点和连续的时间内进行描述的一个高级的概念. 物理边界描述了镜头, 语义边界则描述了场景. 镜头是指使用单个镜头进行连续拍摄的片段, 且视频序列内容也没有明显变化, 是一段视频序列的基本组成单元, 镜头边界检测 (shot boundary detection)[7] 是指将视频片段分割到镜头层面的处理操作. 帧是构成完整运动画面的静止图像之一, 是视频中的最小单位. 关键帧是由于连续帧之间的相似性, 因此需要根据镜头内容的复杂性从单个镜头中选择一个或多个关键帧, 所选择的关键帧即代表着当前视频帧的内容.

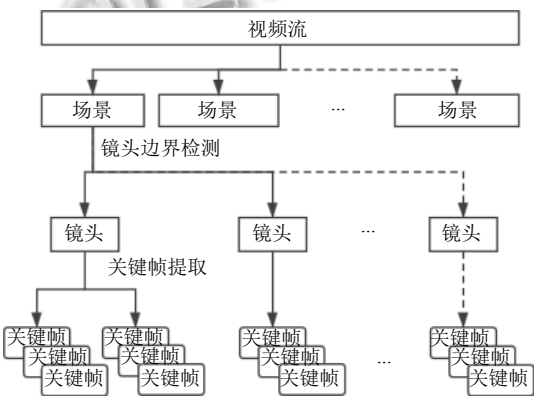

图 2 视频分层结构

2 基于语义的视频检索系统结构

基于语义的视频检索系统的总体结构如图 3 所示. 包括如下几个部分: 结构分析, 包括镜头边界检测、关键帧提取和场景分割; 特征提取, 即从视频图像中提取

特征; 视频挖掘, 即对提取到的特征进行挖掘; 视频标注, 即对提取特征的语义索引的构建和对相关知识的挖掘; 用户查询, 即在视频数据库中搜索所需的视频; 相关性反馈, 即通过相关性反馈优化搜索结果.

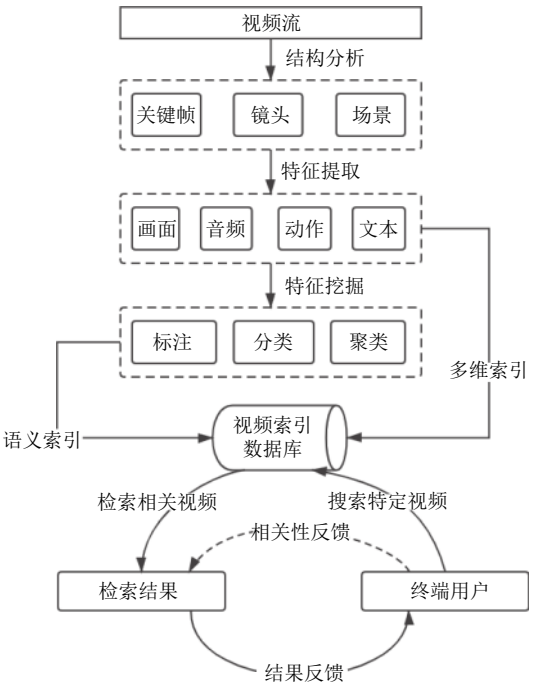

图 3 基于语义的视频检索系统结构

2.1 结构分析

首先通过镜头检测算法将视频分割成多个镜头, 然后确定能够代表该镜头的关键帧.

镜头边界检测是指将整个视频流分割成多个镜头, 在镜头边界位置的帧与其下一帧在视觉特征上是相当不同的, 这是大多数镜头检测算法所依赖的基本原则. 镜头边界指的是连续镜头突变或渐变 (如溶解、淡入、淡出、擦除等) 的转折点[8]. 镜头边界检测常用的方法有: 阈值法[9] 将帧与帧之间的相似性与预先设定的阈值进行比较; 统计法将镜头的边界检测作为分类任务, 可以采用支持向量机 (Support Vector Machines, SVM) 的监督学习算法[10] 和模糊 K-means (Fuzzy K-means) 的无监督学习算法[11] 等方法进行分类.

由于同一镜头的帧存在冗余, 因此选择一个或者多个最能反映镜头内容的帧作为关键帧来表示镜头, 提取关键帧的关键在于选择最能反映镜头内容同时尽可能避免冗余的帧[12]. 可以利用颜色直方图、边缘图和低层形状特征等方式确定关键帧, 关键帧的提取可以基于顺序比较[13]、参考帧[14]、聚类算法[15] 和对象-事件模式[16] 等.

2.2 特征提取

特征是视频数据中的描述性参数, 视频数据的特征描述一般分为: 低层特征、高层特征、对象特征和运动特征等.

低层特征可以从关键帧中提取, 包括从完整图像中提取的全局特征和所选图像部分的局部特征. 颜色特征的典型表示包括颜色直方图、颜色矩阵和颜色相干向量等, 其中使用最多的是颜色直方图, 它描述了图像中每种颜色的相对数量. 纹理特征可以通过 Gabor 滤波器[17]、小波变换[18]、方向特征[19] 和共现矩阵[20] 等方式来提取. 形状特征可以通过连接物体的边缘线, 从关键帧的物体的轮廓中提取. 边缘直方图描述符 (Edge Histogram Descriptor, EHD)[21] 是一种用于边缘检测的算法, 使用直方图描述边缘的分布.

对象特征包括对象所在区域内的颜色、形状和纹理等特征, 可以根据这些相关特征来返回可能包含相似对象的视频片段. 对象表示法是一种描述对象的方法, 通过该方法可以方便地从视频流中检测和检索出对象. 一般可以用物体的形状来表示, 例如基于原始的几何形状、轮廓和边界线, 也可以用物体的外观来表示. 对象特征的缺点是视频中对对象的识别比较复杂, 目前还是主要专注于识别对象的特定部分, 比如仅针对手部等.

运动是动态视频的基本特征, 它携带了视频的时间信息, 与颜色、纹理等其他特征相比, 更接近于客观的语义概念. 基于运动的特征分为两类: 第 1 类是基于相机镜头的运动特征, 例如放大缩小、向左向右平移、向上向下倾斜等; 第 2 类是基于物体本身的运动特征. 运动统计法[22], 视频帧中的点在视频中形成运动分布图, 从而提取统计运动的特征. 运动轨迹法[23], 通过对视频中物体运动轨迹的建模, 提取轨迹特征, 这些特征的准确性依赖于运动视频中正确的分割和目标跟踪. 对象关系法[24], 对多个对象之间的关系进行描述, 而这些特征的缺点是很难标记每个对象及其位置.

视频中的文本是对视频进行自动标注和建立索引的关键信息, 帧或帧序列中的文本会根据其不同的属性展示不同的变化, 如运动状态、颜色状态、几何状态以及边缘状态等. 由于文本区域对噪声比较敏感, 在分辨率较低时, 需要对文本特征进行增强处理, 同时可

以采用光学字符识别 (Optical Character Recognition, OCR) 技术提取文本特征并将其转换为纯文本.

2.3 视频挖掘

视频挖掘是从视频数据中挖掘发现特定的匹配模式及其相关性, 从而提取出未被发现的内容的过程.

视频的语义事件是人们在观看视频时能够理解的高层次语义信息, 视频事件的检测技术试图使计算机对事件的感知能力接近于人类对事件的感知能力. 而导致计算机对视频事件理解困难的原因有很多, 例如目标检测和跟踪的不准确、某些事件的画面发生变化、不同事件的画面表现相似、事件语义的定义解释存在歧义等.

使用无监督或半监督学习技术来自动检测未知的匹配模式, 利用匹配模式可以检测挖掘出与当前匹配模式不同的非寻常事件. 匹配模式挖掘还可以发现一些特殊的内容, 例如挖掘相似的运动模式[25] 和挖掘相似的目标对象[26].

视频关联挖掘可以定义为检测不同事件之间的未知关系, 识别不同对象之间的关联模式的过程.

2.4 视频标注

在基于语义的视频检索中, 视频标注是为视频镜头分配语义概念的过程, 如人、车、天空和行人等. 视频标注和视频分类的一个不同之处在于视频分类一般适用于整个视频, 而视频标注通常使用的是视频镜头作为基础组成单元. 由于视频标注技术有助于弥合语义鸿沟, 因此它也是视频分析任务的基础, 自动化生成视频标注至今仍然是一项艰巨的任务. 基于学习技术, 视频标注可以分为 3 类: 监督学习[27] 需要足够数量的标记训练样本来学习每个概念的具有鲁棒性的检测器, 并且需要的数量随着特征维数的增加而急剧增加; 主动学习[28] 是将无标记样本与监督学习技术相结合来解决无标记样本问题的一种有效方法; 半辅助学习[29] 也是一种利用未标记样本增加已标记样本信息的有效方法.

2.5 用户查询

视频检索的目的是返回用户查询的最相关的视频, 而不同的提交查询数据会得到非常不同的查询结果.

查询类型可以分类为基于非语义的查询, 例如按对象查询和按示例查询等, 以及基于语义的查询, 例如按关键字查询和按自然语言查询等. 按示例查询, 用户提供一个图像或视频作为示例, 以便在该查询中检索所需的视频. 从特定的图像或视频示例中提取低层特征, 然后通过特征相似性度量确定相似视频; 按草图查询, 视频草图由用户绘制, 以便使用它们检索所需的视频; 按对象查询, 利用用户提供的对象图像, 在系统视频数据库中检索出现的所有该对象; 按关键字查询, 用一组关键字描述用户的查询, 它能够从视频中获得一定程度的语义信息; 按概念查询, 也称为语义查询, 它是关键字查询和示例查询的扩展, 用以缩小查询结果范围, 它依赖于具有与视频内容信息相关概念的语义标注; 按自然语言查询, 这是表示语言查询中最自然也是最合适的方向, 这种类型查询的难点在于分析和从自然语言中派生出正确的语义信息; 基于组合的查询, 集成各种类型的查询, 如关键字查询和对象查询, 它适用于多模型的系统.

根据用户对检索系统的查询提交, 将相似度度量技术应用于数据库中的视频检索. 一些常见的相似性度量依据包括欧氏距离 (Euclidean distance)、平方弦距离 (squared chord distance)、卡方距离 (chi-squared distance)、发散度和相关性等. 根据查询类型, 选择用于度量视频相似性的方法. 特征匹配方法[30] 根据对应帧的特征之间的距离来度量视频与查询条件之间的相似度. 文本匹配方法[31] 采用归一化处理后的向量空间模型来计算概念描述文本与查询文本之间的相似性. 组合匹配方法[32] 结合不同的匹配方法, 它能够适应多种模式.

2.6 相关性反馈

相关性反馈将用户查询条件带入系统循环检索,用以缩小提交查询所表示的内容和用户所想内容之间的差距. 相关性反馈是对检索结果的优化, 相关性反馈根据查询条件和返回视频之间的相似性, 对检索到的视频进行评分排名来反映用户所表达意思的优先级. 根据检索结果列出视频, 以便于最相关的视频在检索列表的顶部呈现给用户. 显式相关性反馈[33] 要求用户确定选择相关的视频, 显式反馈因为直接利用了用户的反馈, 所以反馈效果较好, 但也需要更多的互动和用户的配合. 隐式相关性反馈[34] 当用户点击检索到的视频时, 记录此次点击用以优化检索结果, 与显式反馈不同, 隐式反馈不需要用户协作, 更容易被接受和实施, 但从用户处收集的信息不如显式反馈的信息精确. 伪相关性反馈[35] 在没有用户干预的情况下, 从已有的检索结果中选择正样本和负样本, 再将这些样本送回系统中进行研究处理, 虽然伪相关性反馈无需与用户进行交

互, 但语义的理解差距导致伪相关性反馈在应用中受到一定限制.

3 视频检索技术的应用

近年视频检索技术在商业、工业和教育等领域都进行了一定规模的应用, 以下选择主要从视频盗版检测、视频广告监管以及其他方向的应用进行阐述.

3.1 视频盗版检测方向的应用

随着互联网技术的发展, 近年来中国网络核心版权的产业规模迅速增长, 核心版权包括大众所熟知的视频、音乐、文学、游戏、广告以及图片等, 国内视频网站也越来越重视版权价值并将维护版权作为发展重点. 与此产生鲜明对比的是网络视频盗版给企业特别是著作权方带来了严重的损失, 并且这种影响是全球性的, 盗版造成的损失与正版产生的收入呈正相关, 在越来越多正版视频出现的同时, 视频的盗版现象也越来越严重. 典型的侵权模式主要是用户通过下载、破解等手段从拥有正版版权的视频网站上非法下载内容, 经过一些包括添加水印、广告在内的剪辑、加工处理后, 将盗版文件上传至网盘、集合类视频网站等平台供其他用户非法下载观看从而获取不正当收益.

视频检索技术可以实现在大规模的视频数据中检索出近似重复的视频片段, 便于精准、快速打击盗版视频. Chou 等[36] 提出了一种基于时空模式的分层过滤框架下的近重复视频检索与定位方法, 通过基于模式的索引树 (Pattern-based Index Tree, PI-Tree), 快速过滤掉非近似重复的视频, 再设计基于 m 模式的动态规划 (m-Pattern-based Dynamic Programming, mPDP) 算法来定位近似重复的视频片段. da Silva 等[37] 提出了一种相似自连接 (similarity self-join) 的聚类策略, 视频数据集中所有彼此相似的元素进行自连接操作, 将近似重复的视频片段聚集起来进行定位. 当被盗视频被进行一些加工处理, 例如被添加广告水印或被做了剪辑时, 对近似重复视频检索技术便会产生一定的影响造成一定程度的误判. 为提高在视频画面发生变化时检索的准确率, D'Amiano 等[38] 提出了一种用于检测和定位画面发生一些变化的被拷贝视频的方法, 通过快速随机化 Patch 匹配算法和分层分析策略, 对被遮挡、旋转和压缩的近似重复视频片段也具有较好的检测和定位能力.

3.2 视频广告监管方向的应用

视频广告作为数字视频中的一个重要组成部分, 正潜移默化地影响着人们的生活, 其作为商业信息的重要载体, 在传递商业信息上起着无可替代的作用. 随着视频广告数量的不断增加和广告播放方式的多样化, 通过视频检索技术对特定广告进行监管和识别, 有利于支撑广告动态分成业务生态, 轻松把控广告投放的时间、次数等, 同时保障了广告版权方和投放平台的利益, 另外, 基于此技术可以进行广告的高效识别、替换及广告位竞拍.

在海量视频集中对广告商品准确、快速的识别和定位, 有利于平台的广告监管部门对视频中出现的广告进行把控和管理, 可以实现通过广告的分布合理评估营收等应用. Xu 等[39] 提出了一种引入高集成度的多级特征集成模型的方案, 通过更紧密地融合视觉与文本特征信息, 再根据输入的文本数据, 如特定广告物品描述文本, 利用一种双层的长短时记忆 (Long Short-Term Memory, LSTM) 模型直接预测句子查询和视频片段之间的相似度分数, 再使用分段网络过滤掉目标物品不存在的视频片段, 从而可以对出现目标广告物品的视频片段实现定位. Mithun 等[40] 提出了一种多模态视觉线索检索的框架, 根据多模态的视觉线索使用多专家系统 (mixture of expert system) 进行检索. 为了能够更有效地利用视频中可用的多模态线索来完成视频文本检索的任务, 多专家系统注意力主要聚焦于 3 个较为显著和稳定的视频线索, 即物体、活动和地点, 通过对广告商品在这 3 个方面较完整的文本描述, 检索文本与系统模型的组合可以进行较高质量的检索定位工作. 相比直接使用文本进行对广告商品的检索, 当文字概念描述与广告商品本身不容易契合时, 使用商品图片进行检索也是一个可用的选择. Garcia 等[41] 提出了一种基于深度学习 (deep learning) 架构的非对称时空嵌入 (asymmetric spatio-temporal embedding) 模型, 用以在视频集合中根据余弦相似度 (cosine similarity) 找到与输入物品图像最匹配的视频片段. Cheng 等与 Alibaba Group 一同提出了一种新的深度神经网络模型 AsymNet[42], 目标是将视频中出现的商品衣物与线上店铺中相同的商品进行匹配. 从每个视频帧的被检测目标区域中提取深度视觉特征, 并将其输入到 LSTM 框架中进行序列建模, 再对视频的 LSTM 隐藏状态与从静态图像中提取的图像特征进行联合建模,

实现视频中的商品与网上购物图像的精确匹配, 样例效果如图 4 所示, 虚线左边为视频片段, 右边为商品图, 方框圈出部分为匹配结果中细节装饰的差异.

图 4 AsymNet 模型的部分检索匹配结果[42]

3.3 其他方向的应用

视频检索技术除了应用在商业视频领域, 例如视频盗版检测和视频广告监管等方向之外, 还可以应用于城市建设、智能交通、安防监管和教育视讯等领域. 平安城市建设作为全国范围的以视频监控应用为主导, 兼顾城市管理、交通管理和应急指挥等应用的综合体系, 自然成为智能产品和技术应用的重点. 随着感知型摄像机的硬件实力配合云计算的强大算力进入现实应用中, 可以对海量视频数据进行分析以实现基于语义的视频检索应用, 例如高危人员比对、人脸照片检索、全身像检索、车辆视频管控和防区视频管控等智能应用. 随着城市汽车保有量的迅速增长, 交通问题日渐突出, 交通监视控制系统、交通诱导系统和信息采集系统等在交通管理中逐渐发挥越来越大的作用, 视频检索技术运用在交通领域可以实现对包括车牌、车标、车型、车辆颜色和司乘人员等信息进行自动检索, 对各类交通违法事件也可以实现智能监测. 由于公安、司法监所关押人员的特殊性, 安全管理工作尤为重要, 智能视频检索技术用在监狱监所中, 可以实现警戒线检测、剧烈运动检测、起身检测、区域逗留检测、视频遮挡检测等应用, 方便快速发现监所内人员及设备的异常状况, 及时做出处理措施, 有效遏制所内各类突发事件进一步发展. 在教育信息化的大背景下, 传统的现场教学已经无法满足远程教学、后期回看等教学要求. 通过视频检索技术, 可以实现对教师教学细节的跟踪记录, 后期可根据教学场景进行画面切换, 为学生、老师实时或后期观看时提供更好的体验.

4 总结与展望

将视频内容具有的特征转化为人类的语义概念, 是近年来备受关注的研究课题. 本文综述了基于语义的视频检索技术的研究, 视频检索算法的本质任务是根据用户提交的查询, 从给定的数据集合中返回相似的视频, 挖掘和提取视频信息中的语义概念以及如何跨越语义鸿沟的问题仍然是现今视频检索系统中面临的主要挑战. 目前还没有一种完全通用的框架可以用于各种视频的语义特征提取, 当前检索系统的研究应用大多是为了提高特定领域的检索性能和效率. 当系统自动检测语义的特征时, 更精确的检测设备对于检测结果准确率的提高有很大帮助. 相关性反馈通过收集用户在搜索过程中的反馈信息, 是对查询进行迭代更新的有效方法, 查询结果得到改进, 检索性能也会得到提高. 检索模型对检索结果具有决定性的影响, 通过合理的策略组合获得多模态和多概念的学习模型, 可以发挥检索模型和多概念学习模型各自的优势, 提高检索系统的性能. 虽然在视频检索领域已经做了大量的科研工作, 但仍有一些方向可以进一步研究发展:

(1) 分层次解析视频内容画面的特征信息, 以选择合适的特征用于语义概念检测. 视频在不同的层次上通常会包含不同的语义信息, 按照特定的规则提炼不同层次的语义信息, 再针对不同层次的特征使用不同的映射或学习方法, 可以减小单层特征信息交叉解析时带来的影响偏差.

(2) 提升概念探测器的性能, 提高概念检测的速度和精度. 在用户进行查询条件输入时, 可以直接从中提取高级语义概念将其转换生成合适的概念检测器, 对视频片段中的语义概念进行检测, 缩减处理流程. 再通过循环迭代接收相关性反馈信息, 根据反馈不断完善检测方法提升检测精度.

(3) 融合不同的机器学习方法获得更准确的语义概念. 如何提高对广泛概念的识别性能仍然是一个极具挑战性的问题, 尤其是对于较稀有的概念. 近年来通过引入各种不同的机器学习方法, 结合跨模态检索技术对视频片段的语义概念构建准确度对比传统方法有显著的提升, 结合深度学习的检索方式已然成为视频检索领域的热点.

参考文献

1 Snoek CGM, Worring M. Concept-based video retrieval. Foundations and Trends in Information Retrieval, 2008, 2(4): 215–322. [doi: 10.1561/1500000014]

2 Ansari A, Mohammed M. Content based video retrieval systems-methods, techniques, trends and challenges. International Journal of Computer Applications, 2015, 112(7): 13–22.

3 Albanese M, Turaga P, Chellappa R, *et al*. Semantic video content analysis. In: Schonfeld D, CF Shan, DC Tao, *et al*. eds. Video Search and Mining. Berlin, Heidelberg: Springer, 2010. 147–176. [doi: 10.1007/978-3-642-12900-1_6]

4 Hauptmann AG, Christel MG, Yan R. Video retrieval based on semantic concepts. Proceedings of the IEEE, 2008, 96(4): 602–622. [doi: 10.1109/JPROC.2008.916355]

5 Huang X, Shen CY, Boix X, *et al*. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 262–270.

6 Memar S, Affendey LS, Mustapha N, *et al*. Concept-based video retrieval model based on the combination of semantic similarity measures. Proceedings of the 2013 13th International Conference on Intellient Systems Design and Applications. Salangor, Malaysia. 2013. 64–68.

7 Pal G, Rudrapaul D, Acharjee S, *et al*. Video shot boundary detection: A review. In: Satapathy SC, Govardhan A, Raju KS, *et al*. eds. Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. Cham: Springer, 2015. 119–127. [doi: 10.1007/978-3-319-13731-5_14]

8 任会峰. 基于内容的视频检索研究进展. 智慧工厂, 2018, (10): 67–70.

9 Lee H, Yu J, Im Y, *et al*. A unified scheme of shot boundary detection and anchor shot detection in news video story parsing. Multimedia Tools and Applications, 2011, 51(3): 1127–1145. [doi: 10.1007/s11042-010-0462-x]

10 常虹, 张明. 一种基于支持向量机的镜头边界检测算法. 现代计算机, 2016, (20): 73–77. [doi: 10.3969/j.issn.1007-1423.2016.20.015]

11 Lo CC, Wang SJ. Video segmentation using a histogram-based fuzzy c-means clustering algorithm. Computer Standards & Interfaces, 2001, 23(5): 429–438.

12 胡志军, 徐勇. 基于内容的视频检索综述. 计算机科学, 2020, 47(1): 117–123. [doi: 10.11896/jsjkx.190100231]

13 Liu GZ, Zhao JM. Key frame extraction from MPEG video stream. Proceedings of the 2010 3rd International Symposium on Information Processing. Qingdao, China. 2010. 423–427.

14 Sun ZH, Jia KB, Chen HX. Video key frame extraction based on spatial-temporal color distribution. Proceedings of 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Harbin, China. 2008. 196–199. [doi: 10.1109/IIH-MSP.2008.245]

15 Yu XD, Wang L, Tian Q, *et al*. Multilevel video representation with application to keyframe extraction. Proceedings of the 10th International Multimedia Modelling Conference. Brisbane, Australia. 2004. 117–123.

16 Kang HW, Hua XS. To learn representativeness of video frames. Proceedings of the 13th annual ACM International Conference on Multimedia. Singapore. 2005. 423–426.

17 Roslan R, Jamil N. Texture feature extraction using 2-D Gabor filters. Proceedings of 2012 International Symposium on Computer Applications and Industrial Electronics. Kota Kinabalu, Malaysia. 2012. 173–178.

18 Thepade SD, Yadav N. Novel efficient content based video retrieval method using cosine-haar hybrid wavelet transform with energy compaction. Proceedings of 2015 International Conference on Computing Communication Control and Automation. Pune, India. 2015. 615–619. [doi: 10.1109/ICCUBEA.2015.126]

19 赵莹, 高隽, 陈果, 等. 一种基于分形理论的多尺度多方向纹理特征提取方法. 仪器仪表学报, 2008, 29(4): 787–791. [doi: 10.3321/j.issn:0254-3087.2008.04.022]

20 Minarno AE, Munarko Y, Kurniawardhani A, *et al*. Texture feature extraction using co-occurrence matrices of sub-band image for Batik image classification. Proceedings of the 2014 2nd International Conference on Information and Communication Technology. Bandung, Indonesia. 2014. 249–254.

21 Kanagavalli R, Duraiswamy K. Shot detection using genetic edge histogram and object based video retrieval using multiple features. Journal of Computer Science, 2012, 8(8): 1364–1371. [doi: 10.3844/jcssp.2012.1364.1371]

22 Ma YF, Zhang HJ. Motion texture: A new motion based video representation. Proceedings of Object Recognition Supported by User Interaction for Service Robots. Quebec City, QC, Canada. 2002. 548–551.

23 Little JJ, Gu Z. Video retrieval by spatial and temporal structure of trajectories. Proceedings of SPIE 4315, Storage and Retrieval for Media Databases 2001. San Jose, CA, USA. 2001. 545–552. [doi: doi: 10.1117/12.410966]

24 Ngo CW, Pong TC, Zhang HJ. Motion-based video representation for scene change detection. International Journal of Computer Vision, 2002, 50(2): 127–142. [doi: 10.1023/A:1020341931699]

25 Lai C, Rafa T, Nelson DE. Mining motion patterns using color motion map clustering. ACM SIGKDD Explorations Newsletter, 2006, 8(2): 3–10. [doi: 10.1145/1233321.1233322]
26 Anjulan A, Canagarajah N. A novel video mining system. Proceedings of 2007 IEEE International Conference on Image Processing. San Antonio, TX, USA. 2007. I-185–I-188.
27 Snoek CGM, Worring M, van Gemert JC, *et al*. The challenge problem for automated detection of 101 semantic concepts in multimedia. Proceedings of the 14th ACM International Conference on Multimedia. Santa Barbara, CA, USA. 2006. 421–430.
28 Song Y, Hua XS, Dai LR, *et al*. Semi-automatic video annotation based on active learning with multiple complementary predictors. Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval. Singapore. 2005. 97–104.
29 Ewerth R, Freisleben B. Semi-supervised learning for semantic video retrieval. Proceedings of the 6th ACM International Conference on Image and Video Retrieval. Amsterdam, the Netherlands. 2007. 154–161.
30 Browne P, Smeaton AF. Video retrieval using dialogue, keyframe similarity and video objects. Proceedings of IEEE International Conference on Image Processing 2005. Genova, Italy. 2005. III-1208.
31 Snoek CGM, Huurnink B, Hollink L, *et al*. Adding semantics to detectors for video retrieval. IEEE Transactions on multimedia, 2007, 9(5): 975–986. [doi: 10.1109/TMM.2007.900156]
32 Yan R, Yang J, Hauptmann AG. Learning query-class dependent weights in automatic video retrieval. Proceedings of the 12th Annual ACM International Conference on Multimedia. New York, NY, USA. 2004. 548–555.
33 Chen LH, Chin KH, Liao HY. An Integrated Approach to Video Retrieval. Proceedings of the 19th Australasian Database Conference. Wollongong, Australia. 2008. 49–55.
34 Ghosh H, Poornachander P, Mallik A, *et al*. Learning ontology for personalized video retrieval. Proceedings of Workshop on Multimedia Information Retrieval on the Many Faces of Multimedia Semantics. Augsburg, Germany. 2007. 39–46. [doi: 10.1145/1290067.1290075]
35 Jiang L, Yu SI, Meng DY, *et al*. Bridging the ultimate semantic gap: A semantic search engine for internet videos. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. Shanghai, China. 2015. 27–34. [doi: 10.1145/2671188.2749399]
36 Chou CL, Chen HT, Lee SY. Pattern-based near-duplicate video retrieval and localization on Web-scale videos. IEEE Transactions on Multimedia, 2015, 17(3): 382–395. [doi: 10.1109/TMM.2015.2391674]
37 da Silva HB, do Patrocínio ZKG, Gravier G, *et al*. Near-duplicate video detection based on an approximate similarity self-join strategy. Proceedings of the 2016 14th International Workshop on Content-Based Multimedia Indexing. Bucharest, Romania. 2016. 1–6.
38 D'Amiano L, Cozzolino D, Poggi G, *et al*. A patchmatch-based dense-field algorithm for video copy-move detection and localization. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(3): 669–682. [doi: 10.1109/TCSVT.2018.2804768]
39 Xu HJ, He K, Plummer BA, *et al*. Multilevel language and vision integration for text-to-clip retrieval. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, HI, USA. 2019. 9062–9069.
40 Mithun NC, Li JC, Metze F, *et al*. Joint embeddings with multimodal cues for video-text retrieval. International Journal of Multimedia Information Retrieval, 2019, 8(1): 3–18. [doi: 10.1007/s13735-018-00166-3]
41 Garcia N, Vogiatzis G. Asymmetric spatio-temporal embeddings for large-scale image-to-video retrieval. Proceedings of British Machine Vision Conference 2018. Newcastle, UK. 2018. 40.
42 Cheng ZQ, Wu X, Liu Y, *et al*. Video2Shop: Exact matching clothes in videos to online shopping images. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 4169–4177.