

基于内容的视频检索综述

胡志军^{1,2} 徐 勇³

1 贵州大学贵州省公共大数据重点实验室 贵阳 550025

2 贵州大学计算机科学与技术学院 贵阳 550025

3 哈尔滨工业大学(深圳) 广东 深圳 518055

(huzhijuncool@163.com)



摘 要 视频是携带信息量最大的媒体,随着抖音短视频等 APP 的兴起,网络以及数据库的视频数量急剧增加,人工标注的方法已经无法胜任视频检索的任务。视频检索通过提取视频帧的空间特征或者帧与帧之间的时间特征,使得用户能够更客观、更高效地进行视频查找与归类。文中概述了基于内容的视频检索算法,归纳总结了视频检索的一些经典算法,并总结了深度学习在基于内容的视频检索中的研究与应用,最后分析了深度学习在视频检索中的发展前景。

关键词: 视频检索;卷积神经网络;关键帧;特征提取;镜头分割

中图法分类号 TP391

Overview of Content-based Video Retrieval

HU Zhi-jun^{1,2} and XU Yong³

1 Guizhou Provincial Key Laboratory of Public Big Data,Guizhou University,Guiyang 550025,China

2 College of Computer Science & Technology,Guizhou University,Guiyang 550025,China

3 Harbin Institute of Technology(Shenzhen),Shenzhen,Guangdong 518055,China

Abstract Video is the medium with plenty of information,with the rise of short video APP such as vibrato,the number of videos in the network and database has increased dramatically and the method of manual labeling is no longer suitable for video retrieval. Video retrieval by extracting the spatial characteristics of video frames or temporal characteristics between frames and frames enables users to perform video search and categorization more objectively and efficiently. This paper summarized the content-based video retrieval algorithms,some classical algorithms of video retrieval,and the research and application of deep learning in content-based video retrieval. Finally,the development prospect of deep learning in video retrieval was analyzed.

Keywords Video retrieval,Convolutional neural network,Key frame,Feature extraction,Shot segmentation

1 引言

随着 4G 网络的普及,5G 网络的研发,智能手机、数码相机等硬件设备的升级,以及微信、抖音等 APP 和 Facebook、新浪等知名网站上短视频功能的开通,网络视频业务蓬勃发展。YouTube 公布了一项数据,该数据统计了一个月内使用移动设备在 YouTube 上观看视频的时间超过 1h 的用户数量,2013 年 6 月的用户数为 10 亿,而 2017 年 6 月该数据变为 15 亿,增长速度非常快^[1]。2018 年 6 月,随着智能手机的普及,我国有 6.09 亿网民观看网络视频,比 6 个月前多了 3 014 万,其中 5.78 亿为手机视频用户,比 6 个月前多了 2 929 万,5.94 亿的短视频用户占有网民的 74%。由于利益的驱使,各大网络平台加大了对网络短视频的投入,《2018 年中国网络视听传播发展报告》对 2018 年视频内容行业规模给出了惊人的预测结果,行业规模高达 2 016.8 亿元,相比 2017 年增长

39.1%^[2]。面对如此多的视频数据,如何有效地从这些视频库中检索出人们感兴趣的视频,已经成为当今信息化时代的一个难题。因此,近年来为了解决这个难题,视频检索(CB-VR)应运而生。视频检索是指输入一个视频片段并在数据库找出一个或者多个与该输入视频相似的视频并返回给用户。与图像相比,视频提供了各种复杂的视觉模式,包括每帧中的低层视觉内容以及跨帧的高层语义内容,这使得视频检索比图像检索更具挑战性。视频检索的关键在于视频特征提取和视频特征相似度测量技术。

相比传统的依靠手工标注的基于文本的视频检索,基于内容的视频检索(CBVR)能够更高效、更客观地完成视频检索任务,因为视频数据库中的视频数量已经显著增加,人工标注的方法已经无法处理拥有海量视频的数据库。现有的视频检索系统基于颜色、大小、形状、纹理等底层特征进行视频检索,但检索的精度和效率较低。为了能够判断视频的语义内

到稿日期:2019-01-28 返修日期:2019-06-13 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:贵州省公共大数据重点实验室开放课题基金(2018BDKFJJ001)

This work was supported by the Foundation of Guizhou Provincial Key Laboratory of Public Big Data (2018BDKFJJ001).

通信作者:徐勇(laterfall@hit.edu.cn)

容,必须集成低层次和高层次的特征^[3]。随着深度学习技术的不断发展,深度学习在视频分类和视频检索方面受到了广泛的关注,其不仅在学习底层特征时有着良好的性能,在高层特征的学习上也有卓越的表现^[4]。

本文第2节将详细综述传统视频检索技术的概念以及实现步骤;第3节介绍了深度学习在视频检索方面的应用;最后总结全文并对视频检索的发展趋势进行了展望。

2 视频检索技术

2.1 视频检索概念

视频数据按语义概念可抽象表示为4层,自上而下分别是视频层、场景层、镜头以及视频帧^[5]。视频层由一系列的图片组成,这些图片配有相应的音频和文本,形成了动画流。场景层可以理解为语义上的相似性,由一个或多个相邻镜头组成。摄像机不经过切换即可形成一个镜头,镜头又可以被分解为连续的帧,相邻的帧在视觉内容上几乎完全相同,即相邻帧的特征几乎相同,若相邻图像帧之间的特征发生了明显的变化,则认为发生了镜头变换。一个视频帧可以看作是一幅静止的图象,相邻的视频帧在视觉上的相似性表明可以从一个镜头中提取一个或多个关键帧来代表一个镜头。视频层、场景层、镜头以及关键帧之间的关系如图1所示。

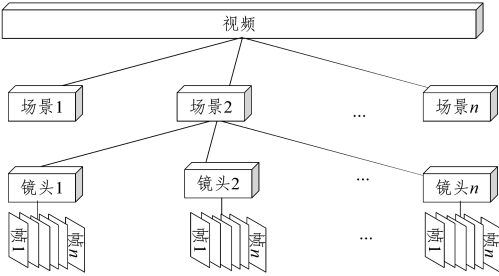


图1 视频的层次结构
Fig. 1 Structure of video

一个基于内容的视频检索(CBVR)系统通过视频内容来帮助操作员检索大型数据库中与查询视频相似的视频^[6],其方法是对数据库中的每一个视频进行镜头检测、关键帧提取、特征提取、索引分类、特征入库等操作,最终形成按索引号分类的特征库,在检索时对查询视频进行镜头检测、关键帧提取、特征提取、建立索引和相似度匹配等操作。从图像到视频的移动给检索问题增加了复杂性,这是由于相比于图像,视频需要在时序方面进行索引、分析和浏览^[7]。视频检索的流程如图2所示。

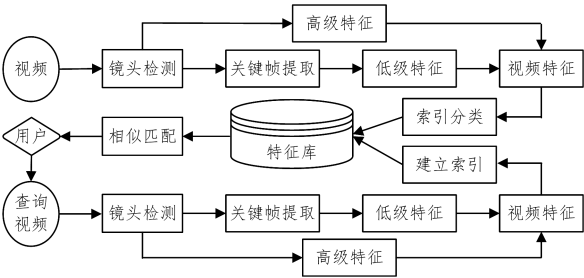


图2 视频检索流程图
Fig. 2 Flow chart of video retrieval

(空间特征)和高级语义(时间特征)之间的鸿沟。通常情况下,纹理和颜色等低级特征是全局的特征,容易被理解和提取,其计算过程也相对简单,然而将低级特征与语义相连接是一个巨大的挑战,尤其涉及到人类的智力和情感时。

2.2 视频镜头检测

镜头检测用于将整个视频分割成多个镜头。镜头边界位置处的帧与属于下一镜头的连续帧之间存在视觉差异,这是大多数镜头边界检测算法所依赖的基本原理。镜头的边界分为突变和渐变,突变是连续镜头之间的突然过渡,而渐变是软过渡,软过渡可以是不同类型的,如硬切、裁剪、溶解、擦拭和淡入淡出^[8]。通常采用镜头边界检测的方法有:像素比较法、颜色直方图作差法、统计学习方法和深度学习方法。

像素比较法也称作模块匹配法,是估计连续两个帧之间相应像素亮度和色度的差值的方法。其最简单的方法是对相邻帧之间的像素差值求和,如果和大于一个预定的阈值,则说明两个帧之间存在镜头边界^[9-10](镜头切换)。由于上述方法对噪声和相机运动非常敏感,文献^[11]在对相邻帧像素做差之前,先用 3×3 的滤波器对帧进行平滑,从而降低噪声的干扰和摄像机运动的影响。

直方图比较法是比较连续两帧直方图的差值,如果所有直方图差值的绝对值之和大于一个阈值,则认为它有一个镜头切换^[12]。文献^[13-14]采用RGB颜色空间进行颜色量化,运用 L_1 范数表示距离。文献^[15]采用HSV空间进行颜色量化,忽略V分量,仅用H分量和S分量建立颜色直方图。文献^[16]对视频帧进行分块,先将对应块间的3个通道的颜色直方图差异求和,再对所有块的颜色直方图差异和再求和,所得结果即为两帧之间的距离。颜色直方图做差法对旋转、平移和噪声具有鲁棒性,借助归一化,颜色直方图还对尺度变换具有不变性。其缺点是忽略了颜色的空间分布信息,两幅完全不同的图片可能有完全相同的颜色直方图。

统计方法的思想是将镜头检测作为分类任务,如采用SVM等有监督学习和模糊c-均值聚类等无监督学习。具体方法是先得到视频帧的特征表示,再用分类器进行分类。常虹等^[17]先求出视频帧颜色直方图和LBP纹理特征,再用这两个特征构造帧间差的特征向量,并用SVM分类器将帧间差分类为有镜头转换或者没有镜头转换。文献^[18]提取颜色直方图得到帧间差直方图向量,并用聚类方法将帧间差分为有镜头转换类和无镜头转换类。聚类算法虽然消除了镜头边界检测对阈值的依赖性,但其缺点是运行时间长,而且需要提前设定聚类的类别数,这是一个难题。

深度学习法是近年来开始被重视的监督学习方法,已有学者开始将其应用于镜头边界检测。Gygli^[19]创建了具有一百万帧的视频帧训练集,通过训练集可以端对端地输出有镜头转换帧与没有镜头转换帧,测试时其检测速度是传统方法的120倍。相比Gygli设计的训练集Hassanien等^[20]设计了更大的视频帧集,超过485万帧,由CNN得到帧间差特征并送入SVM分类器以分配标签。深度学习方法虽然训练速度较慢,但如果训练集选择得好,可以将训练出的参数用于所有边界检测场合,测试时有速度快、准确率高等优点。

2.3 关键帧提取

由于同一镜头的帧存在冗余,因此选择一个或者多个最

CBVR的关键问题之一是弥合语义鸿沟^[7],即低级特征

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。关键帧的提取方式可以分为人工指定、参考帧、聚类、运动能量分布和深度学习方法等。

人工指定关键帧是指不需要提取特征,也不需要帧之间进行比较,而直接由人为指派的关键帧,其主要方法有把每个镜头的第一帧用作关键帧^[21]、对镜头帧序列等间隔取样^[4]和选择每一个镜头的中间帧或首尾两帧^[22]等。这种提取方法不能提供视频镜头的足够信息,特别是对于变化很大的镜头,因此一般很少采用人工指定的方法,这里不做详细介绍。

参考帧的方法是先生成一个参考帧,然后在镜头中选择与该参考帧最相近的帧作为关键帧。Ferman 等^[23]构造了一个阿尔法修剪的平均直方图作为参考帧。SUN 等^[24]用同一个镜头中所有帧的同一个像素位置 (i, j) 处出现次数最多的像素值作为参考帧的像素值。该方法的优点是对存在大量摄像机运动或多个物体运动的视频具有鲁棒性,但其缺点是每次只能从镜头中取出一个帧,当镜头跨时较长时,一帧图像不能很好地表示一个镜头。

聚类的方法是类似于参考帧的方法,首先对一个镜头的所有帧进行聚类,然后选择距离中心最近的帧作为关键帧。文献[25]利用颜色特征子空间中的模糊 K 均值聚类提取关键帧。聚类的一个缺点是聚类的簇数需要预先设置,然而一个镜头中的关键帧个数是根据镜头的复杂程度决定的,文献[25]的方法无法达到自适应选择关键帧个数的要求,而文献[26]提出的自适应阈值聚类方法弥补了这个缺点。基于聚类算法的优点是能够使用通用的聚类算法,并且所提取的关键帧能够反映视频的全局特征,但其缺点是依赖于聚类结果,但聚类的结果和实际语义有一定的差距,而且视频的时间特性不能自然地利用,时间上相差很远的帧有可能被分配到同一个簇。

人工指定和统计方法都没有考虑视频中物体的运动,如当视频中运动员打羽毛球时,他击球时扬起手臂的图片比站立休息时的图片作为关键帧更具有说服力。Wolf^[27]利用帧的光流的水平分量和垂直分量得到动量函数,选择局部极小值作为关键帧,该方法可以在同一个镜头中找出多个关键帧。Liu 等^[28]开发了感知运动能量三角形模型,选择三角形顶点处的帧作为关键帧。Ejaz 等^[29]提出了一种将视觉注意机制、相对运动强度和相对运动方向 3 个特征相融合的关键帧提取方案。

深度学习也可用于关键帧的提取,Hoang 等^[30]提出了一种基于快速深度神经网络 SegNet 的视频摘要方法来学习和识别视频序列中的关键帧,但该方法的缺点是只能得到视频的近似帧,而且不能自适应地选择视频镜头的关键帧数量。Yan 等^[31]设计了一个由摘要器和鉴别器组成的卷积网络。摘要器是一个双流卷积网络,用于捕获视频帧的外观和运动特征,并将其编码成视频特征;鉴别器用于区分视频帧是否为关键帧。深度学习是一种监督学习方法,需要大量的有标签的训练集,深度学习提取关键帧的缺点是对关键帧设置标签时掺杂了太多的主观因素。

由于人们关注的重点不同,导致关键帧的提取具有很大的主观性,关键帧选择时用到的特征不同,提取出的关键帧也

就不同,因此目前还没有一个好的标准来衡量关键帧提取的好坏。

2.4 特征提取

镜头分割和关键帧提取都需要提取特征,本节所讨论的特征提取主要是指易于进行视频索引并最终用于视频相似度计算的视觉特征。这些特征可以来源于视频中提取的关键帧,或者视频中提取的物体,或者视频的运动特征。通常,视频数据特征的性质可以分为低级特征和高级特征。

低级特征也叫空间特征或静止特征,是基于单帧的特征,这里通常以关键帧的视觉特征来代替镜头的视觉特征,因此低级视觉特征忽略了关键帧与其相邻帧之间的时间关系,但低级视觉特征的提取比较简单,通常可以全自动地进行提取和索引。最常用的低级视觉特征包含颜色特征、纹理特征和形状特征。

颜色特征是一种最重要的低级视觉特征,最常用的颜色空间有 RGB 颜色空间、YcrCb 颜色空间和 HSV 颜色空间。Chun 等^[32]使用 HSV 颜色空间中的色调和饱和度分量图像的颜色自相关图作为颜色特征来进行视频检索。Lin 等^[33]提取了两组视觉特征,其中第一组为 YCbCr 颜色直方图和自相关图,第二组为 RGB 颜色直方图和颜色矩。Cheung 等^[34]利用 HSV 颜色直方图来表示视频剪辑的关键帧,并设计用于视频相似性检测的视频签名聚类算法。基于颜色的特征具有反映人类视觉感知,提取简单,计算复杂度低,对旋转、平移和各种变形具有鲁棒性等优点。其缺点是忽略了颜色的空间分布信息,由于镜头分割是在同一视频中相邻帧之间的比较,因此缺少空间分布信息时对镜头分割的影响不大,但本文是将大型数据库视频中的所有关键帧进行对比,因此影响较大。

视频纹理特征是视频关键帧某邻域内像素灰度级或者颜色的某种变换,包含关于物体表面的组织以及它们与周围环境的相关性的关键信息。常用的纹理特征包括方向特征、基于小波变换的纹理特征、共生矩阵等。Chun 等^[32]采用值分量图像的 BDIP 和 BVLC 矩作为其纹理特征,在高分辨率小波域中提取颜色和纹理特征并进行融合形成视频特征。文献[35]对于 TRECVID-2003 视频检索任务,在使用颜色直方图和运动向量直方图的同时,还使用了粗糙度、对比度和方向性等纹理特征。Lin 等^[33]提取的两组视觉特征中,也用到了纹理特征,第一组用到的纹理特征有边缘方向直方图、Dudani 不变矩和共生矩阵,第二组用到的纹理特征有粗糙度、方向等。基于纹理的特征的优点在于它们能够有效地反映视频信息,而且计算复杂度低,然而这些特征在非纹理视频图像中是不可用的。

视频帧的形状特征也是描述视频内容的一个重要特征,主要方法有基于曲率的方法和基于边缘的方法。Dyana 等^[36]用帧间差和混合高斯模型来提取前景对象,并提取关键帧的前景对象的形状,利用形状曲率尺度空间得到形状特征,最后与该前景对象的质心的运动轨迹特征相结合,形成视频特征进行视频检索。Potluri 等^[37]运用中心像素与周围 8 个像素的差的绝对值的和来提取关键帧的边缘特征,从而进行视频检索。Foley 等^[38]首先将图像分割成 4×4 块,然后提取每个块的边缘直方图来进行视频检索。基于形状的特征对

于形状信息在视频中突出的应用是有效的,但是它们比基于颜色或纹理的特征更难提取。

仅用一个低级特征往往很难达到较好的效果, Jiang 等^[39]将多个特征相融合,利用局部关键点的特征词袋模型(BoF)来提取视频帧的 SIFT 特征,并对 SIFT 特征进行聚类,再结合颜色特征和纹理特征进行视频检索,从而取得了很好的效果。该方法表明局部 BoF 特征和全局的颜色、纹理等特征具有较高的互补性。

高级特征也称为时间特征或者运动特征,是区别动态视频和静止图像的基本特征,它比静态关键帧特征更接近视频语义概念。视频检索的运动特征主要是指基于对象的运动特征,其研究方法可以分为基于运动分割的研究方法和基于轨迹的研究方法。

基于运动分割的研究方法是指使用相邻帧之间的差异将图像分割成与不同物体相对应的区域,使用光流方法^[40-41]来估计像素级的运动矢量,然后将像素聚类成相干运动的区域以获得分割结果。使用光流方法进行分割主要有两个缺点:1)光流法不能很好地处理较大的运动;2)相干运动区域可能包含多个对象,需要进一步分割以进行对象提取。文献^[42]针对这些缺点进行了改进,将空间信息结合到运动分割中,对第一帧进行空间分割以获得初始分割结果,然后使用仿射区域匹配来分割后续帧。利用区域分割方法提取特征的优点在于其计算复杂度低。其局限性在于提取的特征不能准确地表示对象行为,也不能刻画对象之间的关系。

物体的运动轨迹也可以作为一种重要的运动特征,通过描述物体的运动轨迹可获取运动特征。文献^[43]使用 SIFT 算子生成点轨迹,通过这些点轨迹分析物体的运动和物体的空间属性,再对这些点轨迹聚类形成运动段,然后基于最常见的 SIFT 对应关系建立估计的运动段之间的时间对应关系。Hsieh 等^[44]提出了一种基于草图和基于句法的混合方案。基于草图的方案是先采用采样技术从每个轨迹中提取一些特征点,再对其进行曲线拟合以测量两个轨迹之间的视距;除了视距外,基于句法的方案使用句法含义来比较任意两个轨迹的距离。Jung 等^[45]基于多项式曲线拟合建立运动模型,运动模型用作访问单个对象的索引。Lai 等^[46]提出了一种以轨迹输入的方式在数据库中查询视频的方法。对于输入的轨迹,系统不仅会返回相似的轨迹,而且会返回形成该轨迹的物体形状。轨迹特征可以描述对象的动作,通过比较动作的轨迹曲线可以比较视频的相似度,但运动轨迹的捕捉需要建立在对目标的正确分割和对物体轨迹的跟踪的基础上,由于同一个动作的运动速度和方向等因素都会影响相似度量,因此轨迹跟踪问题一直是一个具有挑战性的问题。

大部分学者都将低级特征和高级特征进行融合,从而形成视频特征进行视频检索。Kumar 等^[47]采用 HSV 颜色直方图和 MPEG 格式中提取的 p 帧以编译运动直方图进行特征融合。Brindha 等^[48]提取 YcbCr 颜色模型中的色调和饱和度,比较梯度值和阈值来绘制边缘形成纹理特征,并提取形状特征,将这 3 类特征融合成低级特征向量,再用 SVM 分类器对低级特征进行分类,若 SVM 分类器将关键帧标记为匹配实例,则提取深度图像形成运动特征,用 ESN 神经网络对运动特征进行匹配。由于特征之间的互补性,选择合适的特征

进行多特征融合比选择单一的特征进行视频检索能达到更好的检索精度。

2.5 视频索引

由于视频内容的复杂性,所提取的特征通常是高维向量,基于该高维向量得出最终视频检索需要的相似性度量,即用欧氏距离或者余弦距离等距离公式在视频库中寻找与查询视频最相似的视频。若采用全搜索,则对于视频数量非常庞大的视频库非常耗时,此时需要对视频特征库构建索引结构。常用的高维视频数据库特征索引技术包括树结构索引和哈希索引。

树型索引是指将特征向量按照维度分层次构造树结构的索引方式,常用的树型结构有二叉树、X-tree 和 R-tree。二叉树索引结构中最常见的是 kd-tree, R-tree 是一种完全树,当 kd-tree 和 R-tree 的维数较高时,容易产生维数灾难,当维数很高时,查询效率甚至低于全搜索法。文献^[49]采用 kd-tree 为视频帧建立索引,在视频库中用 K-NN 搜索算法搜索出与查询视频的每个关键帧相似度最高的 3 个视频,用这些视频中重复度最高的视频作为查询视频的最佳查询结果。Duan 等^[50]运用空间和时间特征得到视频的特征向量之后,考虑到 kd-tree 在建立索引时存在维数灾难,因此选择 mrkd-tree 建立索引并采用 K-NN 查询方式,在该索引结构下,用户可以根据查询结果决定是否需要进行进一步的搜索。该方法的缺点是只能查询到近似最佳的结果,由于该方法的查询质量损失不大,在需要实时查询结果的情况下,运用该方法可以大大提高检索速度。

视频哈希索引源自图象哈希算法,其将高维的视频特征通过哈希函数映射成固定长度的二进制哈希序列,两个距离相近的视频高维特征向量经过哈希函数被映射到同一个哈希值的概率远远大于两个不相近的视频高维特征向量,两个相同或相近的哈希值代表着相近的视频序列。视频哈希方法可以分为监督哈希算法和无监督哈希算法,它们的区别是在构造哈希函数的过程中是否需要带标签的训练视频集。

无监督哈希算法是指训练过程中的训练视频不需要训练标签就可以得到哈希函数,最常用的是局部敏感哈希算法(Local Sensitive Hash, LSH)。文献^[51]对视频的关键帧提取颜色直方图特征并用 LSH 算法映射成哈希值。Roover 等^[52]提出了一种比直方图向量更稳健的图像像素径向投影算法来提取镜头关键帧的径向方差(Radial Variance, RAV)特征向量,并将 RAV 特征向量的 40 个低频系数量化成 8 位二进制值,最后用该二进制值作为镜头的哈希码。由于文献^[52]的方法没有考虑视频的时间信息, Coskun 等^[53]采用三维余弦变换(3D-DCT)和三维随机基底变换(3D-RBT)对视频镜头进行处理,选择特定的变换系数后采用与文献^[52]不同的哈希函数得到最终的哈希值。文献^[54]用张量分解方法把全局特征、局部特征和时间特征进行融合,从而得到视频特征,其中全局特征采用灰度直方图,局部特征采用 SURF 特征,而时间特征采用相邻视频帧的像素差,然后采用曼哈顿哈希方法得到视频特征的哈希值,并采用十进制的曼哈顿距离来计算不同视频哈希值的距离。传统的哈希方法的二值映射采取的是阈值形式,对高维特征的每一个维度采用同一个阈值进行映射,如果阈值刚好处于该维度数值比较集中的地方,

则会影响映射的精确度,而曼哈顿哈希对每个维度进行了不同的划分,并且不限于二值划分,因此很好地解决了传统哈希的这个缺点。曼哈顿哈希的相似度匹配精度比 LSH 高,但效率不如 LSH,因为曼哈顿采用十进制进行相似度匹配计算。对此,Chen 等^[55]提出了一种加速曼哈顿哈希算法,使用重新映射的方法将曼哈顿哈希码重新映射成二进制数以提高匹配效率。

与无监督视频哈希索引算法相反,监督视频哈希算法在训练过程中需要使用标注信息,并利用标注信息来优化哈希函数,训练视频集的标注准确度影响着训练得到的哈希函数的优劣程度。监督哈希算法常常需要大量的训练集,因此在训练阶段速度很慢,但检索精度较高。常用的无监督哈希算法是深度学习方法。Liong 等^[56]提出了一种 DVH 的深度网络模型,该模型可以端对端地输出视频的哈希码表示,将一组连续帧按顺序传递到卷积层和池化层,从而获得每个帧的 CNN 表示,然后在全连接层中进行融合以得到时间特征,输出层输出二进制表示的哈希码。视频深度哈希方法不仅可以得到更好的检索精度,而且是端对端的输出,消除了中间的特征融合和哈希函数构造阶段,相比传统哈希算法更加高效。但其缺点是需要大量的训练集,并需要对训练集进行标注,这给研究人员带来了不小的挑战。

3 基于深度学习的视频检索

2012 年,Krizhevsky^[57]在 Pascal 竞赛上用卷积神经网络(CNN)对 ImageNet 数据库进行分类,获得了第一名,其 Top5 的错误率仅仅为 15.3%,远远低于第二名的 26%,之后 CNN 在图像分类和图像检索(查找文献)领域得到了广泛的应用^[58],在视频检索领域也有部分研究人员用深度学习的方法对视频检索进行研究。基于深度学习的视频检索方法可以分为带哈希层的端到端方法和不带哈希层的方法。

不带哈希层的深度学习框架是直接从深度网络中学习视频的视频特征,用该视频特征进行视频相似度计算。Kordopatis 等^[59]在可选的已经训练好的 AlexNet, VGGNet 和 GoogleNet 基础之上,提出了一种以提取中间卷积层后面的最大池化层的每个通道的最大值作为视频帧的特征向量的分量值的方法,其采用向量聚合和层聚合两种可选方案形成视频帧特征,对 10 万个随机视频帧的特征向量采用 K-mean++ 聚类算法得到 K 个词袋模型,将视频的关键帧分配到最近的单词中形成视频直方图,即视频特征。Podlesnaya 等^[60]认为已经训练好的 GoogleNet 输出的 1024 维向量已经包含了足够的语义信息,并用其进行镜头分割和镜头相似度匹配,用欧氏距离计算相邻帧的 1024 维特征向量的距离进行镜头检测(若该欧氏距离超过给定阈值则认为是有镜头转换),用余弦距离对查询视频和数据库视频的关键帧的 1024 维特征向量计算相似度,通过搜索最小距离并对最小距离排序进行视频检索。不带哈希层的深度学习方法虽然没有将哈希码集成到深度学习网络框架之中,但它继承了深度网络的优点,即在提取特征时避免了传统非深度学习方法在提取特征时掺杂太多主观因素的缺点,提取出的特征能够很好地表征视频。

带哈希层的深度学习方法是指在设计深度学习框架时,加入一层哈希层,其输出全是二进制码 0 或者 1,即视频的哈

希码。文献[61]在训练阶段,首先在视频中随机地抽取 3 组具有相同帧数量的视频帧,从深度卷积神经网络中提取各输入帧的统一特征表示;然后利用加权平均将各输入视频帧的特征融合为视频特征以简化网络,接着经过两个全连接层,第一个全连接层后面跟着一个 sigmoid 层,该 sigmoid 层用于学习相似性的二进制哈希码,而第二个全连接层具有 k 个节点,其中 k 是类别的数量,即输出层;最后用分类损失和三重损失相结合的损失函数进行优化。文献[62]提出了一种端到端的可以同时学习视频特征和哈希表示的深度视频哈希框架,该框架利用类内特征的同一性和类间特征的多样性来训练视频哈希表示,并过滤掉具有负作用的比特位。文献[61-62]只利用了视频的空间信息。而没有融合视频的时间信息,Gu 等^[63]提出了一种监督递归哈希(SRH),利用深度神经网络进行视频哈希学习,该网络将卷积神经网络学得的原始视频帧的特征输入长短时记忆网络(LSTM),并将 LSTM 的输出作为输入,分别输入到最大池化层和全连接层从而得到两个输出,将这两个输出组合并生成哈希码,最后由损失函数来优化哈希码。Zhang 等^[64]提出了一种无监督哈希框架 SSTH,其以端到端的方式学习视频的时间性质。SSTH 是一个配备了二进制的 LSTM(BLSTM)的循环网络,用视频的帧顺序来自我监督地学习视频的二进制哈希码。带哈希层的深度学习方法可以端到端地生成视频哈希码,最后在视频检索时只需要在与输入视频具有相同或者相近的哈希码的视频集中搜索最相近的视频即可。由于传统哈希码生成过程中的生成函数需要人为设定,即把某些值映射成 0,把另外的值映射成 1,当分界点设置在该维度数值比较集中的地方时,会影响映射的精度,而带有哈希层的深度学习框架是通过损失函数优化这一分界点,可以较好地避免分界点选择不佳的这一缺陷,因此带哈希层的深度学习框架在避免更多的主观因素的同时,提高了哈希码的质量。

结束语 本文介绍了视频检索的概念,并对镜头分割、关键帧提取、特征提取和视频索引所用到的算法进行了详细阐述,介绍了各种算法优缺点。随着互联网的发展和手机硬件的不断更新,视频检索势必会受到更多的关注。在未来视频检索还可以从以下几个方面进行提升:

- (1)可以在交叉模式下进行视频检索方面的研究,如可以运用声音、文本和视觉等特征联合起来进行检索。
- (2)运用用户反馈技术,用户在检索过程中可以与系统进行互动和相互干预,以达到用户期望的检索效果。
- (3)重点发展深度学习在视频检索方面的应用,视频分类是先学习视频特征再学习分类器,而视频检索是先学习视频特征再学习相似度,它们之间的共同点都是需要学习视频特征,而很多学者已经在视频分类方面做出了很大的贡献,可以将这些技术迁移到视频检索中。

参 考 文 献

[1] 网络视听生态圈. YouTube 宣布每月用户已达 15 亿移动视频正在抢电视用户[DB/OL]. (2017-6-23)[2019-1-8]. http://www.sohu.com/a/151603041_728306.
[2] 新传考研小小班. 2018 热词| 视听传播[DB/OL]. (2018-12-5)[2019-1-8]. http://www.sohu.com/a/279939987_736480.
[3] MEGRHIS,SOUIDENE W,BEGHDADIA. Spatio-temporal sa-

- lient Feature extraction for Perceptual Content Based Video Retrieval[C]//IEEE 2013 Colour and Visual Computing Symposium (CVCS). Gjovik, Norway, 2013: 1-7.
- [4] ZOLFAGHARI M, SINGH K, BROX T. ECO: Efficient Convolutional Network for Online Video Understanding[J]. arXiv: 1804.09066, 2018.
 - [5] PAL G, RUDRAPAUL D, ACHARJEE S, et al. Video shot boundary detection: a review[C]//Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2. India: Springer, Cham, 2015: 119-127.
 - [6] MARCHAND-MAILLET S. Content-based video retrieval: An overview [OL]. <https://archive-ouverte.unige.ch/unige:48023>.
 - [7] SEBEN, LEW M S, ZHOU X, et al. The state of the art in image and video retrieval[C]//International Conference on Image and Video Retrieval. Springer, Berlin, Heidelberg, 2003: 1-8.
 - [8] YUAN J, WANG H, XIAO L, et al. A formal study of shot boundary detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(2): 168-186.
 - [9] KIKUKAWA T, KAWAFUCHI S. Development of an automatic summary editing system for the audio visual resources[J]. Transactions of the Institute of Electronics Information & Communication Engineers A, 1992, 75(43): 204-212.
 - [10] LEE M S, YANG Y M, LEE S W. Automatic video parsing using shot boundary detection and camera operation analysis[J]. Pattern Recognition, 2001, 34(3): 711-719.
 - [11] ZHANG H J, KANKANHALLI A, SMOLIAR S W. Automatic partitioning of full-motion video[J]. Multimedia Systems, 1993, 1(1): 10-28.
 - [12] NAGASAKA A, TANAKA Y. Automatic scene-change detection method for video works[C]//2nd Working Conference on Visual Database Systems. Japan Information Processing Society, 1991: 119-133.
 - [13] KWEON I S, HAN S, YOON K. A new technique for shot detection and key frames selection in histogram space[C]//Proceedings of the 12th Workshop on Image Processing and Image Understanding. Korea, 2000: 475-479.
 - [14] YEO B L, LIU B. Rapid scene analysis on compressed video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1995, 5(6): 533-544.
 - [15] QIN J P, FU M S, TU Z Z, et al. Video shot boundary detection based on histogram change ratio[J]. Computer Applications and Software, 2011, 28(4): 17-20.
 - [16] KO K C, CHEON Y M, KIM G Y, et al. Video shot boundary detection algorithm[M]//Computer Vision, Graphics and Image Processing. Springer, Berlin, Heidelberg, 2006: 388-396.
 - [17] CHANG H, ZHANG M. An algorithm of video Sshotboundary detection based on SVM[J]. Graphic and Image, 2016, 7(20): 73-77.
 - [18] LO C C, WANG S J. Video segmentation using a histogram-based fuzzy c-means clustering algorithm[J]. Computer Standards & Interfaces, 2001, 23(5): 429-438.
 - [19] GYGLI M. Ridiculously Fast Shot Boundary Detection with Fully Convolutional Neural Networks[C]//2018 International Conference on Content Based Multimedia Indexing, CBMI 2018. La Rochelle, France, 2018: 1-4.
 - [20] HASSANIEN A, ELGHARIB M, SELIM A, et al. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks[J]. arXiv: 1705.03281, 2017.
 - [21] LI Y, LEE S H, YE H C H, et al. Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques[J]. IEEE Signal Processing Magazine, 2006, 23(2): 79-89.
 - [22] WANG X J, DING H T, CHEN H X. A shot clustering based approach for scene segmentation[J]. Chinese Journal of Image and Graphics, 2007, 12(12): 2127-2130.
 - [23] FERMAN A, TEKALP A. Two-stage hierarchical video summary extraction to match low-level user browsing preferences[J]. IEEE Transactions on Multimedia, 2003, 5(2): 244-256.
 - [24] SUN Z, JIA K, CHEN H. Video key frame extraction based on spatial-temporal color distribution[C]//International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, 2008: 196-199.
 - [25] YU X D, WANG L, TIAN Q, et al. Multilevel video representation with application to keyframe extraction[C]//Proceedings 10th International Multimedia Modelling Conference. IEEE, 2004: 117-123.
 - [26] ZHUANG Y, RUI Y, HUANG T S, et al. Adaptive key frame extraction using unsupervised clustering[C]//Proceedings 1998 International Conference on Image Processing (ICIP98). IEEE, 1998: 866-870.
 - [27] WOLF W. Key frame selection by motion analysis[C]//IEEE International Conference on Acoustics, Speech, & Signal Processing. 1996: 1228-1231.
 - [28] LIU T, ZHANG H J, QI F. A novel video key-frame-extraction algorithm based on perceived motion energy model[J]. IEEE transactions on Circuits and Systems for Video Technology, 2003, 13(10): 1006-1013.
 - [29] EJAZ N, BAIK S W, MAJEED H, et al. Multi-scale contrast and relative motion-based key frame extraction[J]. EURASIP Journal on Image and Video Processing, 2018, 2018(1): 40.
 - [30] HOANG N N, LEE G S, KIM S H, et al. A Real-time Multimodal Hand Gesture Recognition via 3D Convolutional Neural Network and Key Frame Extraction[C]//Proceedings of the 2018 International Conference on Machine Learning and Machine Intelligence. ACM, 2018: 32-37.
 - [31] YAN X, GILANI S Z, QIN H, et al. Deep Keyframe Detection in Human Action Videos[J]. arXiv: 1804.10021, 2018.
 - [32] CHUN Y D, KIM N C, JANG I H. Content-based image retrieval using multiresolution color and texture features[J]. IEEE Transactions on Multimedia, 2008, 10(6): 1073-1084.
 - [33] LIN C Y, TSENG B L, NAPHADE M, et al. VideoAL: a novel end-to-end MPEG-7 video automatic labeling system[C]//In IEEE Intl. Conf. on Image Processing (ICIP). IEEE, 2003, 3: III-53.
 - [34] CHEUNG S C S, ZAKHOR A. Video similarity detection with video signature clustering[C]//International Conference on Image Processing, 2001. Thessaloniki, Greece: IEEE, 2001: 649-652.
 - [35] AMIR A, BERG M, CHANG S F, et al. IBM research TRECVID-2003 video retrieval system[OL]. <https://www.docin.com/p-1550931773.html>.
 - [36] DYANA A, SUBRAMANIAN M P, DAS S. Combining features for shape and motion trajectory of video objects for efficient content based video retrieval[C]//2009 Seventh International Con-

- ference on Advances in Pattern Recognition, Kolkata, India: IEEE,2009:113-116.
- [37] POTLURI T, SRAVANI T, RAMAKRISHNA B, et al. Content-Based Video Retrieval Using Dominant Color and Shape Feature[C]//Proceedings of the First International Conference on Computational Intelligence and Informatics, Springer, Singapore, 2017:373-380.
- [38] FOLEY C, GURRIN C, JONES G J F, et al. TREC'Vid 2005 experiments at dublin city university[OL]. <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>.
- [39] JIANG Y G, NGO C W, YANG J. Towards optimal bag-of-features for object categorization and semantic video retrieval[C]//Proceedings of the 6th ACM International Conference on Image and Video Retrieval, New York, NY, USA: ACM, 2007: 494-501.
- [40] HORN B K P, SCHUNCK B G. Determining optical flow[J]. Artificial Intelligence, 1981, 17(1/2/3):185-203.
- [41] ZHONG D, CHANG S F. Spatio-temporal video search using the object based video representation[C]//Proceedings of International Conference on Image Processing, Santa Barbara, CA, USA: IEEE, 1997, 1: 21-24.
- [42] DENG Y, MUKHERJEE D, MANJUNATH B S. NeTra-V: Toward an object-based video representation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1998, 8(5):616-627.
- [43] BASHARAT A, ZHAI Y, SHAH M. Content based video matching using spatiotemporal volumes[J]. Computer Vision and Image Understanding, 2008, 110(3):360-377.
- [44] HSIEH J W, YU S L, CHEN Y S. Motion-based video retrieval by trajectory matching[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(3):396-409.
- [45] JUNG Y K, LEE K W, HO Y S. Content-based event retrieval using semantic scene interpretation for automated traffic surveillance[J]. IEEE Transactions on Intelligent Transportation Systems, 2001, 2(3):151-163.
- [46] LAI Y H, YANG C K. Video object retrieval by trajectory and appearance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(6):1026-1037.
- [47] KUMAR G S N, REDDY V S K, KUMAR S S. High-Performance Video Retrieval Based on Spatio-Temporal Features[M]//Microelectronics, Electromagnetics and Telecommunications, Springer, Singapore, 2018:433-441.
- [48] BRINDHA N, VISALAKSHI P. Bridging semantic gap between high-level and low-level features in content-based video retrieval using multi-stage ESN-SVM classifier [J]. Sādhanā, 2017, 42(1):1-10.
- [49] FENG Z H, ZHU Y B, LI W Q. Video near-duplicate retrieval based on deep learning[J]. Computer Applications and Software, 2018, 35(1):160-163.
- [50] DUAN L Y, YUAN J, TIAN Q, et al. Fast and robust video clip search using index structure[C]//Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA: ACM, 2004:756-757.
- [51] FERMAN A M, TEKALP A M, MEHROTRA R. Robust color histogram descriptors for video segment retrieval and identification[J]. IEEE Transactions on Image Processing, 2002, 11(5): 497-508.
- [52] DE ROOVER C, DE VLEESCHOUWER C, LEFEBVRE F, et al. Robust video hashing based on radial projections of key frames[J]. IEEE Transactions on Signal processing, 2005, 53(10):4020-4037.
- [53] COSKUN B, SANKUR B, MEMON N. Spatio-Temporal Transform Based Video Hashing[J]. IEEE Transactions on Multimedia, 2006, 8(6):1190-1208.
- [54] NIE X S, WANG S T, YIN Y L. Video hash learning based on feature fusion and Manhattan quantization[J]. Journal of Nanjing University, 2016, 52(4):705-713.
- [55] CHEN W, DING G, LIN Z, et al. Accelerated Manhattan hashing via bit-remapping with location information[J]. Multimedia Tools and Applications, 2017, 76(2):2441-2466.
- [56] LIONG V E, LU J, TAN Y P, et al. Deep video hashing[J]. IEEE Transactions on Multimedia, 2017, 19(6):1209-1219.
- [57] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems 25 (NIPS 2012), Nevada, 2012:1097-1105.
- [58] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016:770-778.
- [59] KORDOPATIS G, PAPADOPOULOS S, PATRAS I, et al. Near-duplicate video retrieval by aggregating intermediate cnn layers[C]//International Conference on Multimedia Modeling, Springer, Cham, 2017:251-263.
- [60] PODLESNAYA A, PODLESNYY S. Deep learning based semantic video indexing and retrieval[C]//Proceedings of SAI Intelligent Systems Conference, Springer, Cham, 2016:359-372.
- [61] DONG Y, LI J. Video retrieval based on deep convolutional neural network[C]//Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing, New York, NY, USA: ACM, 2018:12-16.
- [62] LIU X, ZHAO L, DING D, et al. Deep Hashing with Category Mask for Fast Video Retrieval[J]. arXiv:1712.08315, 2017.
- [63] GU Y, MA C, YANG J. Supervised recurrent hashing for large scale video retrieval[C]//Proceedings of the 2016 ACM on Multimedia Conference, New York, NY, USA: ACM, 2016:272-276.
- [64] ZHANG H, WANG M, HONG R, et al. Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing[C]//Proceedings of the 24th ACM International Conference on Multimedia, New York, NY, USA: ACM, 2016: 781-790.



HU Zhi-jun, born in 1981, doctoral student, lecturer. His main research interests include fractal image compression, image and video retrieval.



XU Yong, born in 1972, Ph.D, professor, Ph.D supervisor. His main research interests include pattern recognition, biometrics, machine learning and video analysis.