计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170003
54(6):1238-1250,2017

视频拷贝检测方法综述

顾佳伟　赵瑞玮　姜育刚
(复旦大学计算机科学技术学院　上海　201203)
(gujw15@fudan.edu.cn)

Video Copy Detection Method: A Review

Gu Jiawei, Zhao Ruiwei, and Jiang Yugang
(*School of Computer Science, Fudan University, Shanghai* 201203)

Abstract　Currently, there exist large amount of copy videos on the Internet. To identify these videos, researchers have been working on the study of video copy detection methods for a long time. In recent years, a few new video copy detection algorithms have been proposed with the introduction of deep learning. In this article, we provide a review on the existing representative video copy detection methods. We introduce the general framework of video copy detection system as well as the various implementation choices of its components, including feature extraction, indexing, feature matching and time alignment. The discussed approaches include the latest deep learning based methods, mainly the application of deep convolutional neural networks and siamese convolutional neural networks in video copy detection system. Furthermore, we summarize the evaluation criteria used in video copy detection and discuss the performance of some representative methods on five popular datasets. In the end, we envision future directions on this topic.

Key words　video copy detection; feature representation; performance evaluation; dataset; review

摘　要　目前网络上存在着大量的拷贝视频,研究人员长期以来致力于视频拷贝检测技术的研究,特别是近年来随着深度学习方法的引入,又涌现出了一些新颖的检测算法.将对现有代表性的视频拷贝检测方法进行回顾与总结,涵盖视频拷贝检测系统的基本框架与各个主要步骤的不同实现方法,包含视频拷贝检测中的特征提取、建立索引、特征匹配与时间对齐等不同模块.总结的关键技术包括了最新的深度学习方法在其中的应用与取得的突破,主要体现在深度卷积神经网络和双胞胎卷积神经网络方法的应用.此外,还将详细介绍目前常用的5个用于视频拷贝检测评测的数据集及通用的评价标准,并讨论分析一些代表性方法的性能表现.最后,对视频拷贝检测技术未来发展趋势进行展望.

关键词　视频拷贝检测;特征表示;性能评价;数据集;综述

中图法分类号　TP311

随着互联网的快速发展,承载着人类活动信息的网络数据正以指数速度增长.据统计,这些海量的网络数据中80%的内容为图像视频等媒体数据[1].例如,全球最大的视频网站YouTube在2007年初平均每分钟有6h时长的视频被上传;在2010年11月,该数字增加到了35h;在2013年5月,平均每分钟上传视频进一步增至100h;而至2015年7月,这一数字已攀升至400h[2];与此同时,根据2014年4月的统计结果,人们每个月要花费60亿小时的时间在收看YouTube的视频内容上[3].据IDC在

收稿日期:2017-01-03;**修回日期**:2017-03-07

基金项目:国家自然科学基金优秀青年科学基金项目(61622204)

This work was supported by the National Natural Science Foundation of China for Excellent Young Scientists (61622204).

2012 年预测,到 2020 年全世界网络数据规模将达到 40 ZB[4].

互联网的高速发展是一把双刃剑,它在带给人们方便与快捷的同时,也导致了许多问题. 例如,一些盗版商利用网络平台出售盗版视频以获取不正当利益;一些用户与团体借助网络平台恶意传播非法视频以扰乱社会秩序等. 在这样的背景下,多种问题视频在各个视频网站、交友社区、聊天工具等平台中不断传播,危害社会. 由于网络数据规模十分庞大,依靠人力在海量数据中找出拷贝视频是不现实的,视频拷贝检测技术也因此被提出. 该技术的应用场景是,基于已有的源视频,在海量数据中寻找与之相同或近似的拷贝视频. 视频拷贝检测技术除了可以应对上述的版权保护问题[5-7]与非法内容检测问题[8]之外,还可以处理视频监控计数问题[9]、视频推荐问题[10]等. 比如,一些用户希望知道某视频片段在网络流媒体上某个时间段内出现的次数,获取这类信息就需要运用该技术;当前各类视频网站的个性化推荐服务是促进用户体验的重要手段,除了依据文本标签匹配外,联合视觉内容进行视频推送,可以达到更准确的推送效果. 随着人类社会进入移动互联网时代,多媒体信息传播更加便捷化,形式更加复杂化,越来越多的地方需要用到这种技术.

早期的视频拷贝检测技术主要使用各类传统特征进行检测,取得了不错的结果;近几年,随着深度学习方法的引入,涌现了一批新的基于深度网络模型的视频拷贝检测技术,它们相比传统方法取得了更优秀的识别效果. 针对目前发展现状,本文对现有代表性的视频拷贝检测方法进行回顾与总结,借此希望能给当前及未来的相关研究提供一定的参考与帮助.

1 视频拷贝检测技术概述

1.1 视频拷贝检测技术定义

目前关于视频拷贝检测技术的研究已有十多年,视频拷贝检测技术主要针对拷贝视频进行检测,但在同时期还存在几种相近的检测对象[11],如重复视频、近似重复视频[5,12]等. 重复视频即为几乎一模一样的视频,范围较窄;近重复视频,要求语义一致、画面近似,视频来源一般不同;而拷贝视频,要求语义一致、画面近似且视频来源相同. 例如,父母用各自的手机分别记录某时刻孩子的生活,这 2 个视频视为近似重复而不是拷贝;如果母亲对其中一个视频进行后期加工,加入一些卡通元素,则新视频才被视为拷贝视频. 在研究之初,其定义范围较窄,一些研究者认为拷贝检测与近似重复检测有明显的差异[6]. 后来,Basharat 等人[13]建议放宽定义,以适应更广泛的应用;为了获取大众对近似视频的理解,Cherubini 等人[10]还做了网络调查. 虽然目前还未有统一检测对象,但它们所使用的检测方法是共通的[11].

一般地,拷贝视频主要由源视频经过光学变换、几何变换或时间变换等变换方式转化而得,具体有插入图标、模拟录像、尺度改变和画中画等方式[8,14],图 1 展示了部分拷贝方式. 其中,图 1(a)为亮度改变;图 1(b)为左右对称变换;图 1(c)为插入图标;图 1(d)为画中画. 在实际应用中,视频的拷贝变换具有多样性与不确定性,研究者希望找到一些通用的方法来适应所有的拷贝变换,目前许多方法对各种变换都有一定的效果,但在不同变换上存在着一定差异,一般插入图标和改变伽马值等拷贝变换较易检测,而模拟录像、画中画和后期加工等拷贝变换的检测比较困难[8,15-18]. 从图 1 中可以看出,后者在视觉内容上的变化相对较大.

(a) Brightness change (b) Bilateral symmetry transform

(c) Insertion of patterns (d) Picture in picture

Fig. 1 Examples of copied frames

图 1 拷贝帧样例

另外,视频拷贝检测应对不同任务具有不同的检测级别. 一些研究工作仅考虑整个视频是否拷贝[12,19],即对于一个查询视频,在参考集中找出与整个查询视频互为拷贝的视频,这种检测被视为全局视频拷贝检测. 相对地,更细粒度的局部视频拷贝检测技术主要针对视频中的任意片段,找出 2 个视频中所有的拷贝片段对[7,14,20-21]. 局部视频拷贝检测虽然具有更为全面、精准的效果,但检索过程相对复杂,导致了检索效率的降低.

1.2 视频拷贝检测技术基本框架

典型的视频拷贝检测技术基本框架如图 2 所示,它主要包含 4 个步骤:特征提取(feature extraction)、建立索引(indexing)、特征匹配(feature matching)和

时间对齐(temporal alignment).框架中对于数据库视频(database videos)的建模为离线步骤(图 2 中 offline 线路);而对于查询视频(query video)需要进行更复杂的在线检测步骤(图 2 中 online 线路),下面介绍其大致流程.

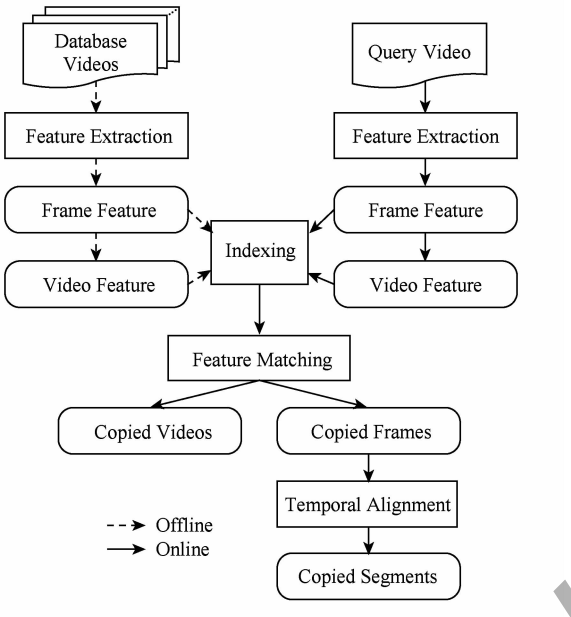

Fig. 2 A general framework of a video copy detection system

图 2 视频拷贝检测技术基本框架

如图 2 所示,无论对于视频库中的视频还是查询视频,首先需要进行特征提取步骤,即对视频关键帧提取相应的特征向量,并经过一定处理形成帧特征或视频特征.具体的特征提取方法将在第 2 节中详细介绍.值得一提的是:一个视频主要由帧序列和音频信息组成,视频拷贝检测技术主要关注其帧序列.音频信息对于拷贝检测的帮助不够稳定,因为相似音频固然能给予加分,但音频上差异较大的拷贝视频反而可能会被误判为非拷贝视频.因此音频信息在视频拷贝检测上并不具备普适性[12],故一般不被采用.

在获得帧特征或视频特征之后,需要进行建立索引操作.对于海量数据库视频中的拷贝检测问题,使用直接的特征一一匹配方式显得十分耗时.为了达到更高效的检索,建立索引是一种必要的手段.理想的索引结构不仅要能提高检索速度,还应控制因建立索引而产生的量化误差.

对于数据库视频,只需执行以上步骤即可.而对于一个查询视频,还需要进行之后的特征匹配操作,也可称之为索引匹配.不同的索引结构有不同的索引匹配方式.针对不同任务,如果是全局视频拷贝检测,通常将大于阈值的匹配结果确认为拷贝视频;如果是局部视频拷贝检测,则将大于阈值的匹配结果确认为拷贝帧.第 3 节介绍了 4 种比较有代表性的索引结构与特征匹配方法.

最后,针对局部视频拷贝检测,还需要使用时间信息把拷贝帧整合成拷贝片段.具体的时间对齐方法详见第 4 节.

1.3 讨论与分析

1.2 节介绍了视频拷贝检测技术的基本框架,特征提取部分对视频生成具有视觉关键信息描述且又易于后续计算的数字序列.建立索引部分主要考虑的是特征匹配的效率问题,是为了实现高效的实时在线检测系统而采用的一种技术.建立索引的同时往往会损失一定精度,对于不同的应用场景,是否使用索引结构以及使用何种索引结构都需要权衡考虑.时间对齐部分主要用于提取 2 个视频的具体拷贝片段,相比全局视频拷贝检测,局部视频拷贝检测具有更直观、更精确的效果,但同时带来低效的检索效率也是不可避免的.

2 特征提取

对视频拷贝检测系统中的特征提取环节,研究者总希望找到一种通用的特征,使之能够鲁棒地应对各种拷贝变换,可以说视觉特征是视频拷贝检测的关键[20-22].

目前对于视频的描述特征分为 2 类:

1) 对于视频帧级别的特征描述,该类特征大量用于局部视频拷贝检测中.在早期的工作中,大量传统的图像特征提取方法被用于视频帧级别的特征提取.近年来,随着深度学习技术的兴起,出现了一些基于深度网络的视频帧特征提取方法.

2) 融合视频内的所有帧信息后的视频整体描述特征,主要用在全局视频拷贝检测问题,它在计算上依赖于前者视频帧级别的特征描述.

以下先回顾常用的基于传统方法和基于深度网络的视频帧特征的提取算法,再对视频全局特征提取方法进行简单介绍.

2.1 基于传统方法的视频帧特征

颜色直方图与尺度不变特征变换(scale-invariant feature transform, SIFT)是视频拷贝检测系统中极为常用的 2 种传统视频帧特征提取方法.

在计算颜色直方图时,需要预设一定的颜色域,对于原始图像的像素矩阵,统计每一个像素点的颜

色值,对其所属的颜色域进行计数,整个方法描述的是不同色彩在整幅图像中所占的比例.由于计算量小、检索高效,该方法及其改进方法被运用于许多相关工作[6,12,23-29].然而颜色直方图只考虑颜色信息,而忽略了视频帧的几何关系、形状信息和纹理信息等,因此具有一定局限性.

SIFT 特征对旋转、尺度缩放、亮度变化保持不变性,对视角变化、仿射变换、噪声也保持一定程度的稳定性[30-32].计算 SIFT 特征时需要对原始图像中的局部关键点进行检测,这些关键点依据各自在原图像上的相对位置而形成几何相关的描述子集合.为了提高匹配效率,研究者采用视觉词袋模型把一个帧内众多的局部描述子合成一个单一特征来表征视频帧,这种特征在视频拷贝检测上具有良好的扩展性和较好的准确率[33].一些研究者针对词袋模型产生的量化误差,使用海明嵌入(Hamming embedding)[34]、基于重叠域的全局上下文描述子(OR-GCD)[35]等方法对其进行了优化.除了词袋模型,一些工作还采用了其他特征编码方式,比如 Fisher Vector 等[36-38].此外,对于视频拷贝检测这一特定任务,有学者还专门提出了对于 SIFT 的改进方法[39-40],例如在文献[39]中,作者结合奇异值分解运算提出了一种名为 SVD-SIFT 的算法.相比于原始的 SIFT 算法,作者指出该改进的特征在保持了尺度、旋转不变性等良好特性的同时,减少了总计算开销,提高了拷贝检测的速度.

2.2 基于深度学习方法的视频帧特征

2012 年 Krizhevsky 等人提出了著名的深度卷积神经网络 AlexNet,它在 ImageNet 挑战赛中的大规模图像分类任务上取得了突破性的成绩[41].此后,大量基于深度学习的方法在计算机视觉领域涌现并取得了巨大成功.在多媒体拷贝检测方面,一些工作[42-43]展示了其远高于传统方法的优异性能.目前深度学习技术在视频拷贝检测方面的成功应用主要集中在使用卷积神经网络和双胞胎卷积神经网络进行视频帧的特征提取.

2.2.1 卷积神经网络方法

卷积神经网络(convolutional neural network, CNN)可以直接用于视频帧特征提取.经典的 AlexNet 主要包含 5 个卷积层(convolutional layer)和 3 个全连接层(fully-connected layer).Jiang 等人[42]采用了预训练的 AlexNet 模型,取 AlexNet 的第 6 层特征(fc6)作为视频帧特征,如表 1 所示.表 1 中的第 1 列是网络各层的名称,第 2 列是对应的输出特征尺寸.该方法使得每一个关键视频帧,都被转化成一个 4 096 维的特征向量.实验表明,该方法具有高于传统方法的优异性能.

Table 1 A Simplified AlexNet Architecture

表 1 简化的 AlexNet 框架

Layer(Low→High)	Output Size
conv1	96×55×55
conv2	256×27×27
conv3	384×13×13
conv4	384×13×13
conv5	256×13×13
fc6	4 096×1×1
fc7	4 096×1×1
fc8	1 000×1×1

在 AlexNet 之后,又有许多深度网络被提出,较著名的有 VGGNet[44], GoogleNet[45] 以及 ResNet[46]等.VGGNet 是一个更深的网络,它最多有 19 层组成,具有更高的辨别能力.同时,它使用更小的卷积过滤器,能够获取原始图像中更多的细节.GoogleNet 包含 22 个网络层,具有多尺度处理能力.一些工作对 GoogleNet 等深度网络框架做了相应的研究,比较了各网络之间的性能差异[47].ResNet 是较新的一个 CNN 框架,它在 2015 年的 ImageNet 挑战赛上获得了冠军.ResNet 使用深度残差网络把 CNN 扩展到了 152 层,而在后续应用中,其深度更是超过了 1 000 层[48].

无论 VGGNet,GoogleNet 还是最近的 ResNet,其网络框架不断变深,这些更先进的网络结构原理上与 AlexNet 相同,都能用在视频帧特征提取.例如,文献[49]的算法基于 VGGNet 进行了视频拷贝检测的相关研究,获得了比 Jiang 等人所提算法更好的结果.

2.2.2 双胞胎卷积神经网络方法

双胞胎卷积神经网络(siamese convolutional neural network, SCNN)[50]由 2 个结构相同、参数共享的子网络组成,它以图像对作为训练输入,通过预测的相似度与实际相似度之间的误差进行前向反馈以调节网络模型参数,如图 3 所示.

该网络用于拷贝检测中的视频特征提取原理是:当网络输入 2 张视频帧图像时,预测的相似度通过欧氏距离计算,模型训练目标是使得拷贝对距离越小、非拷贝对距离越大.SCNN 方法需要准备一定

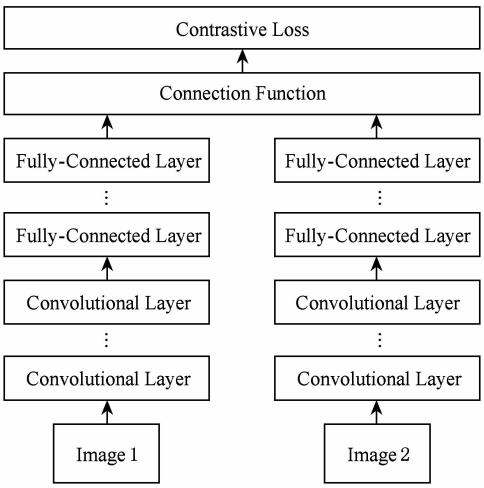

Fig. 3 A basic SCNN architecture

图 3 一个基本的 SCNN 框架

的训练数据,通过模拟拷贝效果的方式制造拷贝对数据,非拷贝对数据可直接抽样随机配对获得.依据所采用的 CNN 框架的不同,视频帧表示方式也有所不同.例如 Jiang 等人在文献[42]中使用了较窄的 CNN 子网络,对视频帧提取多个局部特征,后续计算采用了类似处理 SIFT 特征的方法,即使用词袋模型形成单一向量作为视频帧表示.另一种情况,是使用较宽的 CNN 子网络,可直接提取视频帧的全局特征,如在一些图像拷贝检测的工作[43]采用了此类方法,这些工作也都得到了优于传统方法的良好效果.

SCNN 方法直接针对相似度信息训练模型参数,理论上比标准的深度学习方法更适合多媒体拷贝检测任务.但由于训练数据的差异以及更难的模型训练,SCNN 方法目前取得的整体效果显得并不突出.

2.3 视频全局特征

在视频拷贝检测系统中,除了对以上视频帧描述特征进行比对,还有一类方法将各个视频所有帧的特征合并为全局描述特征,再进行基于视频全局特征的比对.此类方法主要用于全局视频拷贝检测,它的典型代表包括基于视频帧特征聚类的方法[51]、基于视频所有帧特征向量主成分分析得到的边界坐标系统(bounded coordinate system)描述[5,26]、基于视频帧特征直方图统计的累积直方图(accumulative histogram)方法[12]和参考视频直方图(reference video-based histogram)方法[24]等.此类方法的主要优点在于得到的视频描述特征较为精简.然而与基于视频帧特征比较的方法相比,此类方法的最大问题在于它们往往忽略了视频中的局部信息,例如视频片段中出现的物体或区域变化[11].由于视频全局特征的这些不足,近年来提出的一些更有效的拷贝检测系统主要采用基于视频帧比较的方法.

2.4 多特征融合

对于视频拷贝检测问题,大部分已有的方法都只使用一种特征,然而单一特征往往不足以描述视频内容,不能应对复杂而多样的拷贝变换,所以一些方法[19,52]采用了具有不同特性的多重特征作为视频内容描述,获得了比单一特征更好的结果.

3 建立索引与特征匹配

为了达到快速检索的目的,视频拷贝检测系统中通常需要运用高效的索引结构.特别是局部视频拷贝检测,总体特征量十分庞大,如果采用枚举的方式进行一一匹配,检索效率会十分低下,很难应用于在线的视频拷贝检测系统.索引结构一般与特征的形式和特征匹配所采用的最近邻搜索方法相关.本文总结如下 4 种常见的索引方法,分别是树形结构、向量近似文件、Hash 结构和倒排索引方法.

3.1 树形结构

目前已有许多树形索引结构被提出,在视频方面,一种被称为"高斯树"[53]的索引结构既实现了高效率搜索,又保留了较多的视觉信息.高斯树通过管理高斯分布来实现快速的概率查询,它能应用于较复杂的对象.但树形结构对于高维扩展并不友好,当特征维度增加时,会引发"维数灾难问题".

3.2 向量近似文件

向量近似文件(vector approximation file, VA-file)方法[54]的主要思想是将特征空间划分成 $2b$ 个单元,每个单元都可以用一个长度为 b 的二进制比特串表示,查询样本在比对时可以排除距离较远的单元内的数据,从而大大减少了计算开销.后续工作还对算法中不同扫描边界的设定进行了比较分析,提出了 VA-LOW, VA-BND 和 VA-LOW-k 等不同设定以及改进方法[55].

3.3 Hash 结构

Hash 是一种常用的加快查找速度的方法.其中,位置敏感 Hash(locality-sensitive hashing, LSH)[56]能很好地应对高维特征而被广泛应用[40,57].该方法采用一组位置敏感 Hash 函数,在特征空间内做随机方向的线性映射,使得近似的特征能有很高的概率落入同一个散列桶内.LSH 的查询时间是次线性

(sub-linear)的,但同时它的查询结果质量也是不稳定的[58].在随后的几年内,针对其准确率和时空效率,LSH 不断被人改进[59-61].

另外,针对多重特征的情况,多特征 Hash (multiple feature hashing, MFH)[19]被使用于视频拷贝检测并取得了不错的效果.该方法采用一组预训练的 Hash 函数,每个 Hash 函数以多重特征为输入并输出一个二进制位,最后形成一个二进制向量并通过异或操作进行相似度值的计算.MFH 具有较好的扩展性,但其难点在于如何训练 Hash 函数以提高精度与效率.

3.4 倒排索引

倒排索引结构首先被应用于文本检索,后来在图像视频领域也被广泛应用.各类特征通过视觉词袋模型[33,62-63]形成一个个视觉词,所有视觉词形成一个词典,这与文本数据十分相似,因而能很方便地使用倒排索引结构.一个典型的倒排文件主要记录每个视觉词的频率及其出现位置,并以视觉词作为属性、出现位置作为记录,形成属性确定记录的结构.一般地,对于帧级别的特征匹配,以每一帧的局部特征作为属性,整个图像作为记录;对于视频级别的特征匹配,以每一帧的全局特征作为属性,整个视频作为记录.此外,一些研究者针对倒排索引结构造成的几何信息的缺失问题,使用弱几何一致性[34]及其改进方法[9,64]对倒排索引结构进行了优化.

一般地,建立索引会损失一定的量化误差,故为了追求更高的理论精度,一些工作也会不加索引而采用一一匹配的方式[42].特征匹配一般采用距离度量,较普遍的 2 种距离为欧氏距离与余弦距离.针对一些特殊类型的特征,也会采用地球移动距离(earth mover's distance, EMD)[65]、编辑距离(edit distance)[66]等度量方式,其中前者能很好地评估直方图相似性,后者常被运用于类字符串数据.

4 时间对齐

时间对齐是在进行局部视频拷贝检测中为了确定 2 个视频的哪些片段对互为拷贝时进行的操作.对于任意 2 个视频,有一对一、一对多、多对多以及交叉对应等多种拷贝片段对齐形式.图 4 简单描述了以上 4 种情况,图 4 中上下 2 条长线分别表示 2 个完整视频,其中同灰度短线条表示拷贝片段,由中间的指示线连接表明对应关系.为了解决上述形式多样的拷贝片段对齐问题,下面介绍并分析 3 种时间对齐方法,分别是基于滑动窗口的时间对齐算法、基于树形结构的时间对齐算法和基于图的时间对齐算法.

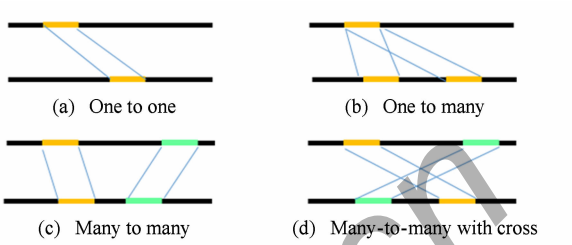

Fig. 4 Four examples of copied segments in a pair of video

图 4 在一对视频内的 4 种拷贝片段对齐情况

4.1 基于滑动窗口的时间对齐算法

Douze 等人[38]提出了一种运用霍夫投票机制的时间对齐方式,该方法首先定义 $s(\tau,t)$ 表示在时刻 τ 的查询帧与在时刻 t 的参考帧的相似度得分.然后分配一个动态窗口,窗口内含有 δ 帧,窗口既可以向右移动,也可以向右扩大.公式 $h(\delta)=\sum_{\tau\in y}s(\tau,\tau+\delta)$ 用于计算滑动窗口变化时的累加相似度直方图,其中 y 是查询视频的时间戳集合,如果时间戳 $\tau+\delta$ 不在参考视频的时间范围内,则 $s(\tau,\tau+\delta)=0$.该直方图具有明显的峰值,通过其峰值确认出拷贝片段.由于某些不相关的匹配帧会形成较高的相似度得分,为了降低这种情况造成的影响,该方法还加入了二次加权策略.

4.2 基于树形结构的时间对齐算法

文献[67]提出了一种树形表示的时间对齐算法.该方法对于每个查询帧,先找出一个与之相似的参考视频帧集合作为候选集,然后将所有候选集构成一个树状结构.树的根节点对应查询视频的某个帧,以该查询帧作为起点,树的第 1 层由其相似帧集合组成,第 2 层由其查询帧的下一帧的相似帧集合组成,并且从第 2 层起,每次连接子树要额外考虑时间信息,即要求在时间上父节点先于子节点且父子节点的时间差小于预设阈值.最后,再使用剪枝策略得到最终的匹配结果.另外,如果构建树的过程中因无法找到可连接的子树而中断时,以中断处的帧为根节点,重新构建树;被中断的树则执行剪枝策略获得相应的拷贝片段.

4.3 基于图的时间对齐算法

Tan 等人[9,68]提出了一种运用网络流算法的时间对齐方式,并开发出了相应的算法工具[69].该方

法对于一个查询视频 Q 和一个参考视频 R，针对视频 Q 中每一帧，从视频 R 中找出与之最相似的 k 帧，用以构建初始的拷贝帧网络；然后严格依据时间顺序，用有向边连接 top-k 列表中的所有帧，边的权重即为对应帧之间的相似度值；最后，执行最大流算法，获得最长的视频拷贝段.

时间对齐是局部视频拷贝检测中的一个重要环节. 上述 3 种方法，第 1 种方法是先考率时间信息，再考虑帧之间的相似度信息；而后 2 种方法与之相反. 基于滑动窗口的方法受视频帧率与预设阈值的影响而不够稳定，基于树与基于图的方法都需要额外确定路径的算法而产生较多的计算量. 针对不同的视频拷贝检测任务，需要使用与之相适应的时间对齐方法.

5 数据集与已有方法性能

5.1 数据集

在视频拷贝检测领域中常见并具有代表性的数据集主要包括 TRECVID[70-72]，Muscle-VCD[7]，CC_Web[12]，UQ_Video[19] 和 VCDB[14]，表 2 中罗列了这些数据集的基本统计信息.

Table 2 Comparison of the Widely Used Copy Detection Datasets[14]

表 2 常见拷贝检测数据集统计信息[14]

Dataset	Year	Partial Copy	Type of Copies
TRECVID2008	2008	Y	Simulated
Muscle-VCD	2007	Y	Simulated
CC_Web	2007	N	Real
UQ_Video	2011	N	Real
VCDB	2014	Y	Real

TRECVID 是美国国家标准技术局(NIST)支持的一个视频检索项目，它在 2008 年发布了一个专用于视频拷贝检测算法评测的公共数据集[72]，该数据集包含 200 h 时长的电视节目视频，约 2 000 个查询片段. 其中，查询片段采样于原数据库，并加以随机的模拟拷贝操作而得，具体操作有插入图标、模拟录像、再编码、后期加工等，一些工作[73-74]涉及了该项目数据集.

Muscle-VCD 数据集包含约 100 h 时长的视频[7]. 该数据集中所有视频采样于网络视频片段、电视档案和电影等，并以不同的比特率、分辨率以及视频格式进行存储. 该评测数据集共有 2 个任务，分别是全局视频拷贝检测与局部视频拷贝检测. 针对不同任务，研究者可以获取一个原始视频集和一个相应的用于查询的模拟拷贝的视频集，以此评估各个视频拷贝检测方法的性能. 一些研究工作[9,75]在该数据集上进行了实验.

CC_Web 是一个包含 12 790 个视频的拷贝检测数据集[12]. 该数据集最大特点是它的所有视频均来源于网络，没有对视频进行模拟拷贝的转化操作，所以这个数据集被认为是体现网络真实拷贝情况的数据集. CC_Web 数据集中的样例视频如图 5 所示. 图 5 中行 a 展示的是原始的视频帧，行 b 的视频帧是经过亮度与尺寸变换后的视频帧，行 c 是调整视频帧采样率后的视频帧，行 d 是加入了文字、分框和内容改变后的结果，行 e-f 是在起始和末尾加入了变化内容的结果，行 g 是整个视频加入上下边框的结果，行 h 是单纯的尺寸变换的结果. 该数据集被运用于一系列工作中[9,12,33,76-78].

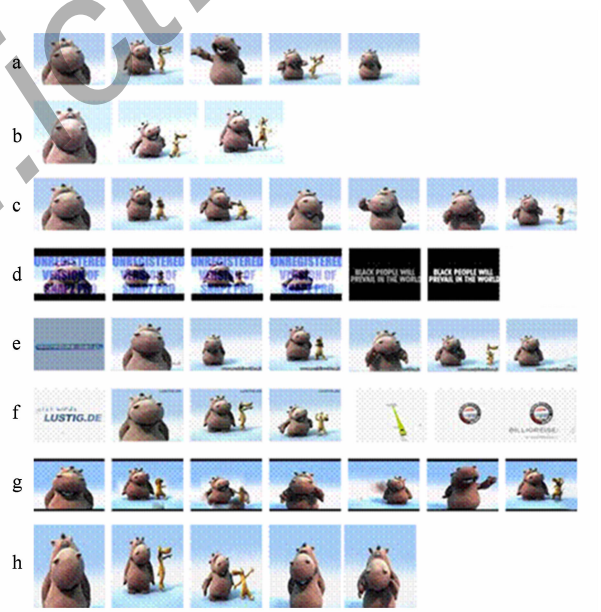

Fig. 5 Examples of video frames in CC_Web dataset[12]

图 5 CC_Web 数据集中视频帧示例[12]

UQ_Video 是对 CC_Web 数据集的扩展，主要加入了 10 多万个干扰视频[19]. CC_Web 和 UQ_Video 这 2 个数据集都只能用于全局视频拷贝检测.

VCDB 是一个较新的视频拷贝检测数据集，它共有 100 528 个视频[14]，其中 528 个视频是核心视频，包含 9 236 对视频拷贝片段，其余的 10 万个视频为干扰视频. 该数据集也完全采集于网络，属于真实拷贝的数据集. 它被用于局部视频拷贝检测，并对各拷贝片段的变化方式做了精确统计. VCDB 数据集中的部分样例视频如图 6 所示. 图 6 中展示了该数

据集中不同类别的视频数据,包括商业、电影、音乐、演讲、运动、监控和其他等主题.图 6 中每张小图的左右两半显示了原始视频帧和经过拷贝变换后的视频帧的样例.基于深度学习的拷贝检测方法[42,49]在该数据集上进行了实验.

Fig. 6 Examples of video frames in VCDB dataset[14]

图 6 VCDB 数据集中视频帧示例[14]

5.2 评价标准

在关于视频拷贝检测方法的评估中,与信息检索相关的准确率、召回率、*F*1 均值以及平均精度均值(mean average precision, *MAP*)都是常用的评测指标.

特别地,在 TRECVID 的拷贝检测任务中,还采用了一种称为最小标准化检测消耗率(minimal normalized detection cost rate, *MinNDCR*)的指标,计算公式如下:

$$NDCR=C_{\text{Miss}}\times P_{\text{Miss}}\times R_{\text{target}}+C_{\text{FA}}\times R_{\text{FA}}, \quad (1)$$

其中,P_{Miss} 与 R_{FA} 分别为漏检率与误检率,C_{Miss} 与 C_{FA} 分别为漏检率与误检率的惩罚系数,R_{target} 为先验达标率.*NDCR* 数值越小,代表检测算法的性能越好.

在 Muscle-VCD-2007 的局部视频拷贝检测任务中,帧精度(QualityFrame, *QF*)和片段精度(QualitySegment, *QS*)指标被用于局部视频拷贝检测算法的性能评测,它们的计算公式分别为

$$QF=1-\frac{|Missed\ Frames|}{|Frames|}, \quad (2)$$

$$QS=\frac{|correct|-|FalseAlarm|}{|Segments|}, \quad (3)$$

其中,*QF* 指标计算的是拷贝片段中帧的覆盖精度,*QS* 指标计算的是拷贝片段的检测精度.*QF* 与 *QS* 的值越大,代表检测算法的性能越好.

在 VCDB 数据集中,视频帧级别的准确率(frame-level precision, *FP*)和召回率(frame-level recall, *FR*)、视频片段级别的准确率(segment-level precision, *SP*)和召回率(segment-level recall, *SR*)指标被用于性能评测,它们的计算公式分别为

$$FP=\frac{|correctly\ retrieved\ \ frames|}{|all\ retrieved\ \ frames|}, \quad (4)$$

$$FR=\frac{|correctly\ retrieved\ \ frames|}{|groundtruth\ copy\ \ frames|}; \quad (5)$$

$$SP=\frac{|correctly\ retrieved\ \ segments|}{|all\ retrieved\ \ segments|}, \quad (6)$$

$$SR=\frac{|correctly\ retrieved\ \ segments|}{|groundtruth\ copy\ \ segments|}, \quad (7)$$

其中,检测返回的一对拷贝段若与实际拷贝段皆有重合,则被视为正确的检索片段(*correctly retrieved segments*).以上指标数值越大,代表检测算法的性能越好.

另外,针对拷贝检测的实际应用场景,一些工作还评测了检测效率[26,33,42,66,79]和可扩展性[9,78,80-81]等指标.

5.3 已有代表性方法性能

目前已有工作的实验对象主要是以上的 5 个数据集,应用场景包括检测拷贝帧、检测拷贝片段以及视频级别检测等.由于各个工作在实验对象、应用场景上的差异,导致无法进行统一的比较.其中 TRECVID 中基于内容的拷贝检测(content based copy detection, CBCD)任务曾提供了一个很好的性能比对平台[71],但这个任务因在 2011 年获得了接近完美的提交结果而被取消了.图 7 展示了 TRECVID 2011 CBCD 中性能最好的前 10 个结果,其中横坐标为拷贝变换方式,纵坐标为 *F*1 得分,数字 1~10 表示排名前 10 的队伍,*Act.* 与 *Opt.* 分别表示为使用队伍提交阈值与使用最优阈值的情况,Median 为所有队伍结果的中位数.图 7 中排名前 10 的队伍采用 *Act.* 阈值的结果用菱形表示,采用 *Opt.* 阈值的结果用短横线表示,Median 结果则用折线图表示,其中折线图上点为方形的为 Act. Median,点为菱形的为 Opt. Median.从图 7 中可以看出,在该数据集上大部分方法都已经达到接近完美的结果.

在 Muscle-VCD 数据集上,表 3 展示了全局视频拷贝检测任务中的部分比较有代表性的算法的评测结果.其中,ADV,IBM,CITYU 和 CAS 是当时 Muscle-VCD-2007 比赛中的拷贝检测方法.从表 3 中可以看出 TNP 方法[9]在精度与效率上都达到了

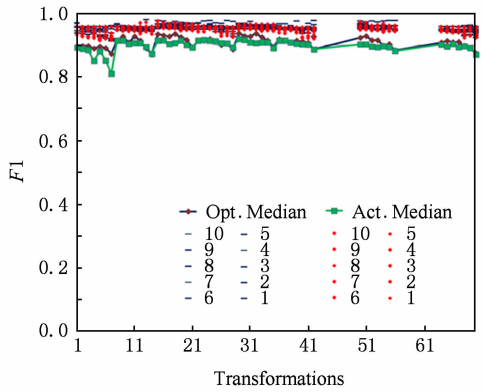

Fig. 7 $F1$ Scores of top 10 performance in TRECVID 2011 Content Based Copy Detection[8]

图 7 TRECVID 2011 CBCD 中 top10 性能的 $F1$ 得分[8]

最优,特别是在精度上获得了 100%的准确率. 表 4 展示了局部视频拷贝检测任务中的部分比较有代表性的算法的评测结果,TNP 方法[9]依旧在精度与效率上展现出最好的性能.

Table 3 The Performance of Several Representative Methods in Entire Video Copy Detection Task of Muscle-VCD-2007

表 3 在 Muscle-VCD-2007 的全局视频拷贝检测任务中部分代表性方法性能

Method	Precision	Runtime/min
TNP[9]	1.00	8.5
Ref[80]	0.93	
ADV	0.86	64
IBM	0.86	44
CITYU	0.66	45
CAS	0.53	15

Table 4 The Performance of Several Representative Methods in Partial Video Copy Detection Task of Muscle-VCD-2007

表 4 在 Muscle-VCD-2007 的局部视频拷贝检测任务中部分代表性方法性能

Method	*QS*	*QF*	Runtime/min
TNP[9]	0.90	0.82	2.85
Ref[80]	0.86		
CITYU	0.86	0.76	35
ADV	0.33	0.17	33

一些工作在 CC_Web 数据集上进行了评测,如表 5 所示. 其中 PPT 方法[21]取得了最优的结果,但差距并不明显,4 种方法都达到了较好的结果;而 SIG_CH 方法[12]由于直接采用颜色直方图所带来的局限性而导致性能不佳.

一些方法在 UQ_Video 数据集上进行了实验,如表 6 所示. 相比 CC_Web 数据集,由于 UQ_Video 数据集增加了 10 多万个干扰视频,各方法性能整体表现不高;其中 PPT 方法[21]依旧取得了最优的结果,但处理速度上不及 MFH 方法[19].

Table 5 The Performance of Several Methods in CC_Web Dataset

表 5 拷贝检测方法在 CC_Web 数据集上的评测性能

Method	MAP
PPT[21]	0.958
MFH[19]	0.954
LBP[78]	0.953
HIRACH[12]	0.952
SIG_CH[12]	0.892

Table 6 The Performance of Several Methods in UQ_Video Dataset[21]

表 6 拷贝检测方法在 UQ_Video 数据集上的评测性能[21]

Method	MAP	Runtime/s
SEQ[75]	0.6663	20 222.54
PI-tree[21]	0.7916	2.88
MFH[19]	0.8618	1.75
PPT[21]	0.8829	5.33

在最新的相关工作中,文献[42]给出了采用深度学习方法与传统方法在拷贝检测任务中的性能比较,如表 7 所示. 实验中,该文作者使用了 VCDB 数据集,并采用 $F1$ 得分作为评价指标. 从该文作者给出的结果与分析中可以明显发现,算法在视频帧级别与视频片段级别上得到的评测结果一致,使用深度学习方法得到的视频特征(CNN 与 SCNN)取得了比传统方法(SIFT)更好的检测性能,验证了深度学习方法在拷贝检测问题中的适用性. 此外,还可以发现同属深度学习方法的 SCNN 网络得到的检测性能并不如普通的 CNN 网络. 文献[42]中作者给出的解释是在使用 CNN 网络时,人们可以利用大量的已标注图像数据对网络进行预训练,而对于 SCNN 可用的训练样本的数据量非常有限,从而影响了最终训练完成的 SCNN 网络的性能.

Table 7 Frame-Level and Segment-Level *F*-Measure on the Core Data Set of VCDB[42]

表 7 不同方法在 VCDB 上的 $F1$ 得分情况[42]

Type	SIFT	CNN	SCNN
Frame Copy	0.6358	0.7101	0.6897
Segment Copy	0.5956	0.6503	0.6317

另外,在局部视频拷贝检测中,时间对齐算法也对检测性能具有一定影响. 文献[14]中将 2 种时间对齐算法在 VCDB 数据集上进行了实验比较,结果表明:在性能上,网络流时间对齐算法在拷贝帧检测与拷贝片段检测上均优于霍夫投票机制时间对齐算法;在效率上,网络流时间对齐算法要略慢于霍夫投票机制时间对齐算法,但在可接受范围内.

6 未来发展趋势

就目前成果而言,拷贝检测技术虽然在一些相对简单的数据集上取得了接近完美的结果,但在相对复杂的数据上还远未达到令人满意的性能. 目前而言,视频拷贝检测方法本身的性能还有待提高,同时更多丰富的评测数据也有待建立,为深入地研究提供帮助.

评测数据集的建立,一方面要考虑其真实性,另一方面要考虑其复杂性. 在真实性上,CC_Web 与 VCDB 等数据集给出了解决方法,即直接从网络环境采集数据;在复杂性上,保证一定量级的同时还要保证拷贝方式的多样化,这个过程中需要一定人为的筛选. 另外,视频标注也是一大挑战,特别是局部视频拷贝检测评测数据集,需要精确拷贝段到秒级;面对如此庞大的标注量,半自动化的标注工具可能是一种解决方法.

在视频拷贝检测方法上,特征表示是其关键. 目前深度学习技术在视频拷贝检测上展现出优于传统方法的性能,这肯定了深度学习的特征表示能力. 未来一段时间内,基于深度学习的视频拷贝检测方法应是主要研究方向,RNN/LSTM 是否可以用来对视频片段建模有待探索;RCNN 这类用于目标检测的网络是否可以用在拷贝检测上也有待研究;更适宜拷贝检测的深度网络结构还有待提出. 与此同时,为了训练出与理论框架效果接近的网络模型,与深度网络结构相适应的训练数据也亟需完善.

另外,随着网络的发展与科技的进步,视频拷贝检测方法所能应对的新的应用场景也将不断被探索.

7 总 结

本文首先描述了视频拷贝检测技术的研究背景;然后介绍了一个实现视频拷贝检测的基本框架,对框架内各步骤要点进行了分析,结合最新的深度学习方法,详细介绍了深度学习在视频拷贝检测方法中的应用与进展;最后回顾了目前具有代表性的 5 个数据集及通用的评价标准,讨论并分析了当前研究状况与未来发展趋势. 随着视频拷贝检测研究的不断深入,希望本文能给当前及未来的研究提供一定的参考与帮助.

参 考 文 献

[1] Analytics Magazine. Images & videos: Really big data [EB/OL]. 2012 [2016-12-05]. http://analytics-magazine.org/images-a-videos-really-big-data

[2] Tubular Insights. 500 hours of video uploaded to YouTube every minute [Forecast] [EB/OL]. 2015 [2016-12-05]. http://tubularinsights.com/hours-minute-uploaded-youtube

[3] Smith G. 145 amazing YouTube statistics (October 2016) [EB/OL]. 2016 [2016-12-05]. http://expandedramblings.com/index.php/youtube-statistics

[4] Infosecurity Magazine. Digital universe is headed for 40 ZB, but big data lacks protection [EB/OL]. 2012[2016-12-05]. http://www.infosecurity-magazine.com/news/digital-universe-is-headed-for-40-zb-but-big-data

[5] Shen Hengtao, Zhou Xiaofang, Huang Zi, et al. UQLIPS: A real-time near-duplicate video clip detection system [C] // Proc of the 33rd Int Conf on Very Large Data Bases. New York: VLDB Endowment, 2007: 1374-1377

[6] Law-To J, Buisson O, Gouet-Brunet V, et al. Robust voting algorithm based on labels of behavior for video copy detection [C] //Proc of the 14th ACM Int Conf on Multimedia. New York: ACM, 2006: 835-844

[7] Law-To J, Joly A, Boujemaa N. Muscle-VCD-2007: A live benchmark for video copy detection [EB/OL]. 2007[2016-12-05]. http://www-rocq.inria.fr/imedia/civr-bench

[8] Kraaij W, Awad G. TRECVID 2011 content based copy detection: Task overview [EB/OL]. Gaithersburg, MD: NIST, 2011 [2016-12-05]. http://www-nlpir.nist.gov/projects/tvpubs/tv11.slides/tv11.ccd.slides.pdf

[9] Tan H K, Ngo C W, Hong R, et al. Scalable detection of partial near-duplicate videos by visual-temporal consistency [C] //Proc of the 17th ACM Int Conf on Multimedia. New York: ACM, 2009: 145-154

[10] Cherubini M, De Oliveira R, Oliver N. Understanding near-duplicate videos: A user-centric approach [C] //Proc of the 17th ACM Int Conf on Multimedia. New York: ACM, 2009: 35-44

[11] Liu Jiajun, Huang Zi, Cai Hongyun, et al. Near-duplicate video retrieval: Current research and future trends [J]. ACM Computing Surveys, 2013, 45(4): No. 44

[12] Wu Xiao, Hauptmann A G, Ngo C W. Practical elimination of near-duplicates from Web video search [C] //Proc of the 15th ACM Int Conf on Multimedia. New York: ACM, 2007: 218-227

[13] Basharat A, Zhai Y, Shah M. Content based video matching using spatiotemporal volumes [J]. Computer Vision and Image Understanding, 2008, 110(3): 360-377

[14] Jiang Yugang, Jiang Yudong, Wang Jiajun. VCDB: A large-scale database for partial copy detection in videos [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2014: 357-371

[15] Jiang Menglin, Fang Shu, Tian Yonghong, et al. PKU-IDM @ TRECVid 2011 CBCD: Content-based copy detection with cascade of multimodal features and temporal pyramid matching [C] //Proc of the TRECVID Workshop. Gaithersburg, MD: NIST, 2011

[16] Ayari M, Delhumeau J, Douze M, et al. Inria @ trecvid'2011: Copy detection & multimedia event detection [C] //Proc of the TRECVID Workshop. Gaithersburg, MD: NIST, 2011

[17] Uchida Y, Takagi K, Sakazawa S. KDDI Labs at TRECVID 2011: Content-based copy detection [C] //Proc of the TRECVID Workshop. Gaithersburg, MD: NIST, 2011

[18] Gupta V, Varcheie P D Z, Gagnon L, et al. CRIM at TRECVID 2011: Content-based copy detection using nearest-neighbor mapping [C] //Proc of the TRECVID Workshop. Gaithersburg, MD: NIST, 2011

[19] Song Jingkuan, Yang Yi, Huang Zi, et al. Multiple feature hashing for real-time large scale near-duplicate video retrieval [C] //Proc of the 19th ACM Int Conf on Multimedia. New York: ACM, 2011: 423-432

[20] Wu Xiao, Li Jintao, Tang Sheng, et al. Video copy detection based on spatio-temporal trajectory behavior feature [J]. Journal of Computer Research and Development, 2010, 47(11): 1871-1877 (in Chinese)
(吴潇, 李锦涛, 唐胜, 等. 基于时空轨迹行为特征的视频拷贝检测方法[J]. 计算机研究与发展, 2010, 47(11): 1871-1877)

[21] Chou C L, Chen H T, Lee S Y. Pattern-based near-duplicate video retrieval and localization on Web-scale videos [J]. IEEE Trans on Multimedia, 2015, 17(3): 382-95

[22] Shinde S, Chiddarwar G. Recent advances in content based video copy detection [C] //Proc of the Int Conf on Pervasive Computing (ICPC 2015). Piscataway, NJ: IEEE, 2015: 1-6

[23] Zobel J, Hoad T C. Detection of video sequences using compact signatures [J]. ACM Trans on Information Systems, 2006, 24(1): 1-50

[24] Liu Lu, Lai Wei, Hua Xiansheng, et al. Video histogram: A novel video signature for efficient Web video duplicate detection [C] //Proc of the 2007 Int Conf on Multimedia Modeling. Berlin: Springer, 2007: 94-103

[25] Wu Xiao, Ngo C W, Hauptmann A G, et al. Real-time near-duplicate elimination for Web video search with content and context [J]. IEEE Trans on Multimedia, 2009, 11(2): 196-207

[26] Huang Zi, Shen Hengtao, Shao Jie, et al. Bounded coordinate system indexing for real-time video clip search [J]. ACM Trans on Information Systems, 2009, 27(3): No. 17

[27] Huang Zi, Hu Bo, Cheng Hong, et al. Mining near-duplicate graph for cluster-based reranking of Web video search results [J]. ACM Trans on Information Systems, 2010, 28(4): No. 22

[28] Jun W, Lee Y, Jun B M. Duplicate video detection for large-scale multimedia [J]. Multimedia Tools and Applications, 2016, 75(23): 15665-15678

[29] Zou Fuhao, Li Xiaowei, Xu Zhihua, et al. Image copy detection with rotation and scaling tolerance [J]. Journal of Computer Research and Development, 2009, 46(8): 1349-1356 (in Chinese)
(邹复好, 李晓威, 许治华, 等. 抗旋转和等比缩放失真的图像拷贝检测技术[J]. 计算机研究与发展, 2009, 46(8): 1349-1356)

[30] Lowe D G. Object recognition from local scale-invariant features [C] //Proc of the 7th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 1999: 1150-1157

[31] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[32] Ke Y, Sukthankar R. PCA-SIFT: A more distinctive representation for local image descriptors [C] //Proc of the 2004 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2004: 506-513

[33] Wu Xiao, Zhao Wanlei, Ngo C W. Near-duplicate keyframe retrieval with visual keywords and semantic context [C] //Proc of the 6th ACM Int Conf on Image and Video Retrieval. New York: ACM, 2007: 162-169

[34] Jégou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search [C] //Proc of the 2008 European Conf on Computer Vision. Berlin: Springer, 2008: 304-317

[35] Zhou Zhili, Wang Yunlong, Wu Q J, et al. Effective and efficient global context verification for image copy detection [J]. IEEE Trans on Information Forensics and Security, 2017, 12(1): 48-63

[36] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization [C] //Proc of the 2007 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2007: 1-8

[37] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation [C] //Proc of the 2010 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 3304-3311

[38] Douze M, Jégou H, Schmid C, et al. Compact video description for copy detection with precise temporal alignment [C] //Proc of the European Conf on Computer Vision. Berlin: Springer, 2010: 522-535

[39] Liu Hong, Lu Hong, Wen Zhaohui, et al. Gradient ordinal signature and fixed-point embedding for efficient near-duplicate video detection [J]. IEEE Trans on Circuits and Systems for Video Technology, 2012, 22(4): 555-66

[40] Zhu Yingying, Huang Xiaoyan, Huang Qiang, et al. Large-scale video copy retrieval with temporal-concentration SIFT [J]. Neurocomputing, 2016, 187(C): 83-91

[41] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C] // Proc of the Advances in Neural Information Processing Systems. New York : Curran Associates, 2012: 1097-1105

[42] Jiang Yugang, Wang Jiajun. Partial copy detection in videos: A benchmark and an evaluation of popular methods [J]. IEEE Trans on Big Data, 2016, 2(1): 32-42

[43] Zhang Jing, Zhu Wenting, Li Bing, et al. Image copy detection based on convolutional neural networks [C] //Proc of the Chinese Conf on Pattern Recognition. Berlin: Springer, 2016: 111-121

[44] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv: 1409.1556, 2014

[45] Szegedy C, Liu Wei, Jia Yangqing, et al. Going deeper with convolutions [C] //Proc of the 2015 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 1-9

[46] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the Conf on the 2016 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770-778

[47] Perkins L N. Convolutional neural networks as feature generators for near-duplicate video detection [R]. Boston, MA: Boston University, 2015

[48] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Identity mappings in deep residual networks [C] //Proc of the 2016 European Conf on Computer Vision. Berlin: Springer, 2016: 630-645

[49] Wang Ling, Bao Yu, Li Haojie, et al. Compact CNN based video representation for efficient video copy detection [C] // Proc of the 2017 Int Conf on Multimedia Modeling. Berlin: Springer, 2017: 576-587

[50] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks [C] //Proc of the 2015 Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4353-4361

[51] Shen Hengtao, Ooi B C, Zhou Xiaofang. Towards effective indexing for very large video sequence database [C] //Proc of the 2005 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2005: 730-741

[52] Lin Ying, Yang Yang, Ling Kang, et al. Video copy detection based on multiple visual features synthesizing [J]. Journal of Image and Graphics, 2013, 18(5): 591-599 (in Chinese)
(林莹, 杨扬, 凌康, 等. 多特征综合的视频拷贝检测[J]. 中国图像图形学报, 2013, 18(5): 591-599)

[53] Bohm C, Gruber M, Kunath P, et al. Prover: Probabilistic video retrieval using the Gauss-tree [C] //Proc of the 23rd Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2007: 1521-1522

[54] Weber R, Schek H J, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces [C] //Proc of the 24th Int Conf on Very Large Data Bases. New York: VLDB Endowment, 1998: 194-205

[55] Weber R, Böhm K. Trading quality for time with nearest-neighbor search [C] //Proc of the 2000 Int Conf on Extending Database Technology: Advances in Database Technology. Berlin: Springer, 2000: 21-35

[56] Datar M, Immorlica N, Indyk P, et al. Locality-sensitive hashing scheme based on p-stable distributions [C] //Proc of the 20th Annual Symp on Computational Geometry. New York: ACM, 2004: 253-262

[57] Xu Zhe, Xue Zhifeng, Chen Fucai. Video copy detection based on improved affinity propagation [J]. Computer Engineering and Design, 2014, 35(9): 3185-3189 (in Chinese)
(许喆, 薛智锋, 陈福才. 基于改进的近邻传播学习算法的视频拷贝检测[J]. 计算机工程与设计, 2014, 35(9): 3185-3189)

[58] Houle M E, Sakuma J. Fast approximate similarity search in extremely high-dimensional data sets [C] //Proc of the 21st Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2005: 619-630

[59] Tao Y, Yi K, Sheng C, et al. Quality and efficiency in high dimensional nearest neighbor search [C] //Proc of 2009 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 563-576

[60] Grauman K. Efficiently searching for similar images [J]. Communications of the ACM, 2010, 53(6): 84-94

[61] Liu Dawei, Yu Zhihua. A computationally efficient algorithm for large scale near-duplicate video detection [C] //Proc of the 2015 Int Conf on Multimedia Modeling. Berlin: Springer, 2015: 481-490

[62] Jégou H, Douze M, Schmid C. Improving bag-of-features for large scale image search [J]. Int Journal of Computer Vision, 2010, 87(3): 316-336

[63] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos [C] //Proc of the 2003 IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2003: 1470-1477

[64] Zhao Wanlei, Ngo C W. Flip-invariant SIFT for copy and object detection [J]. IEEE Trans on Image Processing, 2013, 22(3): 980-991

[65] Jiang Yugang, Ngo C W. Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval [J]. Computer Vision and Image Understanding, 2009, 113(3): 405-414

[66] Huang Zi, Shen Hengtao, Shao Jie, et al. Practical online near-duplicate subsequence detection for continuous video streams [J]. IEEE Trans on Multimedia, 2010, 12(5): 386-398

[67] Can T, Duygulu P. Searching for repeated video sequences [C] //Proc of the 2007 Int workshop on Workshop on Multimedia Information Retrieva. New York: ACM, 2007: 207-216

[68] Tan H K, Ngo C W, Chua T S. Efficient mining of multiple partial near-duplicate alignments by temporal network [J]. IEEE Trans on Circuits and Systems for Video Technology, 2010, 20(11): 1486-1498

[69] Pang Lei, Zhang Wei, Tan H K, et al. VIREO-VH: Video hyperlinking [R]. Hong Kong: City University of Hong Kong, 2012

[70] Smeaton A F, Over P, Kraaij W. Evaluation campaigns and TRECVID [C] //Proc of the 8th ACM Int Workshop on Multimedia Information Retrieval. New York: ACM, 2006: 321-330

[71] Over P, Awad G, Michel M, et al. Trecvid 2011—An overview of the goals, tasks, data, evaluation mechanisms and metrics [C] //Proc of the 2011 TRECVID Workshop. Gaithersburg, MD: NIST, 2011

[72] NIST. Guidelines for the TRECVID 2008 evaluation [EB/OL]. 2008[2016-12-12]. http://www-nlpir. nist. gov/projects/tv2008/tv2008. html

[73] Zhou Xiangmin, Zhou Xiaofang, Chen Lei, et al. An efficient near-duplicate video shot detection method using shot-based interest points [J]. IEEE Trans on Multimedia, 2009, 11(5): 879-891

[74] Douze M, Jégou H, Schmid C. An image-based approach to video copy detection with spatio-temporal post-filtering [J]. IEEE Trans on Multimedia, 2010, 12(4): 257-266

[75] Yeh Mei-Chen, Cheng Kwang-Ting. Video copy detection by fast sequence matching [C] //Proc of the ACM Int Conf on Image and Video Retrieval. New York: ACM, 2009: No. 45

[76] Wu Xiao, Zhao Wanlei, Ngo C W. Efficient near-duplicate keyframe retrieval with visual language models [C] //Proc of 2007 IEEE Int Conf on Multimedia and Expo. Piscataway, NJ: IEEE, 2007: 500-503

[77] Tan H K, Wu Xiao, Ngo C W, et al. Accelerating near-duplicate video matching by combining visual similarity and alignment distortion [C] //Proc of the 16th ACM Int Conf on Multimedia. New York: ACM, 2008: 861-864

[78] Shang Lifeng, Yang Linjun, Wang Fei, et al. Real-time large scale near-duplicate Web video retrieval [C] //Proc of the 18th ACM Int Conf on Multimedia. New York: ACM, 2010: 531-540

[79] Zhao Wanlei, Ngo C W. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection [J]. IEEE Trans on Image Processing, 2009, 18(2): 412-423

[80] Poullot S, Crucianu M, Buisson O. Scalable mining of large video databases using copy detection [C] //Proc of the 16th ACM Int Conf on Multimedia. New York: ACM, 2008: 61-70

[81] Law-To J, Buisson O, Gouet-Brunet V, et al. ViCopT: A robust system for content-based video copy detection in large databases [J]. Multimedia Systems, 2009, 15(6): 337-353

Gu Jiawei, born in 1992. Master candidate of computer science. His main research interests include image and video recognition.

Zhao Ruiwei, born in 1987. PhD candidate of computer science. His main research interests include image and video recognition.

Jiang Yugang, born in 1981. Received his PhD degree in computer science from City University of Hong Kong in 2009. Full professor at the School of Computer Science, Fudan University. Received the NSFC award for outstanding young researchers in 2016. His main research interests include multimedia content analysis and computer vision.