视频相似性检测方法综述

李　雪，王　忠*，范青刚，刘延飞，陈　菁
（火箭军工程大学 基础部，陕西 西安 710025）

摘　要：针对大数据时代服务器端存在众多冗余数据，降低文件检索效率的问题，对代表性的视频相似性检测方法和研究进展进行了总结。首先，介绍了相似性视频的定义及检测流程，分析完整检测系统应具备的视频预处理、特征提取、特征匹配3大模块，并详述了具体实施方法；然后，针对视频相似性检测方法展开阐述，从基于视频时空信息分类、视频特征分类和视频多模态检测方法3个层面深入分析了各方案的实现原理及优缺点；最后，对相似性检测领域常用的数据集进行归纳，对经典算法的检测性能进行了对比，并探讨了视频相似性检测领域存在的问题，对未来发展趋势进行了展望。

关键词：计算机视觉；视频相似性检测；双流特征；卷积神经网络；多模态
中图分类号：TP39　　**文献标志码：**A　　**文章编号：**2097-2741(2024)02-088-015

Review of Video Similarity Detection Methods

LI Xue, WANG Zhong*, FAN Qinggang, LIU Yanfei, CHEN Jing
(Rocket Force University of Engineering, Xi'an 710025, Shaanxi)

ABSTRACT: In the age of big data, numerous redundant data are stored on the server side, which reduces file retrieval efficiency. To solve the problem, the representative video similarity detection methods and research progress were summarized. Firstly, the definition and detection process of similar videos were introduced. Video preprocessing, feature extraction and feature matching, the three necessary modules of a complete detection system were analyzed. Additionally, the implementation of the three modules were introduced in detail. Secondly, video similarity detection methods were elaborated from three levels: video spatiotemporal information classification, video feature classification and video multimodal detection. Moreover, implementation principles, advantages and disadvantages of each scheme were discussed. Finally, commonly used datasets in the field of similarity detection were summarized and sorted out and the detection performance of classical algorithms were compared. Thereby, problems in video similarity detection were discussed and the future development trend was predicted.

KEYWORDS: computer vision; video similarity detection; dual stream features; convolutional neural network; multimodal

随着社交媒体时代信息的爆炸，各类平台每时每刻都在产生大量的、类似的视频数据，平台可以从不断增长的视频中，根据用户的喜好去进行个性化的推荐。海量的视频素材对于用户而言，

收稿日期：2024-02-25
基金项目：火箭军工程大学青年基金（2022QN-S009）
第一作者：李雪（1997—），女，硕士，助教，主要从事云安全、图像处理和机器学习研究。E-mail：1029211440@qq.com
*通信作者：王忠（1968—），男，教授，主要从事嵌入式并行计算与边缘智能研究。E-mail：dsp863wang@163.com

引用格式：李雪，王忠，范青刚，等.视频相似性检测方法综述[J].火箭军工程大学学报，2024，38（2）：88-102.

不仅造成了严重的信息过载，而且每个视频内容的消费者同时也可成为生产者，基于自身兴趣爱好对原始视频进行微小的改动后进行再次存储及转发，这就使得原始视频被重复曝光，为用户带不良体验[1-2]。因此，视频数据量的增长速度不容忽视，对重复数据的检测和删除成为数据存储工作的重中之重。随着机器学习已经成为近些年来的研究热点，卷积神经网络（Convolutional Neural Networks，ConvNet或CNN）[2]由于具有平移不变性、尺度不变性等特征，在图像领域已经取得了较大成功。在很多计算机视觉识别任务中，利用CNN特征的检测效果已经超过了传统的哈希特征提取[3]，因此，如何构造更好的卷积神经网络并将其用于视频去重，是一个重要的研究方向。而视频样本通常包含一系列的视频帧和描述文本。为了更好地评判视频间的相似性特征，部分学者提出可以通过考虑多模态信息来进行衡量，即将视频的语义特征与视频帧特征等信息综合考虑，来学习高质量的多模态表示。故本文基于上述相关背景，对现有的代表性视频相似性检测方法进行回顾和总结，并对该领域的未来研究方向进行展望。

1 视频相似性检测技术概述

目前，针对视频的拷贝攻击方式多种多样，且不同的攻击方式对视频内容所造成的变化也不尽相同，加之大量近乎重复视频的存在，不仅降低了用户体验，还给视频存储、管理、版权保护等方面带来了巨大挑战，因此，快速有效的重复视频检测方案势在必行。

1.1 视频相似性检测技术定义

对于一段视频，如果另一段视频与其在视觉感知内容上基本相同，仅在视频格式、编码参数、光照变化、尺寸等非内容方面的特征不同，则称这2段视频为重复视频[4]，或称后者是前者的拷贝。人眼很容易对相似视频进行判别，但这对于计算机等机器设备而言是十分困难的，主要体现在基于视频内容的改变及非内容的改变等方面[5]。内容改变方面的操作主要包括添加字幕及logo、改变视频背景、画中画等。这些操作对视频添加了额外的信息，但不影响原始视频的语义特征。非内容改变方面的操作主要包括[4]：（1）改变视频格式，如有损压缩、格式扭曲等变化；（2）添加噪声、改变亮度、改变饱和度等变化；（3）视频尺寸的变化，如视频的裁剪、视频帧的几何变化等。

重复数据删除技术作为一种无损的数据缩减方式，已经成为大数据和云计算兴起时代信息数据处理的热门技术，可通过对数据中心或者云端数据的指纹信息进行比对，删除重复数据来达到优化存储资源的效果。此外，视频属于一种非结构化数据，是由多张连续的图片序列堆叠而成，从时间维度上看，视频具有动态性，其特征不仅包括空域特征，还有时间维度上的时域特征。采用不同的特征检测算法，得到的检测结果也不同，且都会关系到检测系统的性能。但不变的是，综合考虑视频的时空域特征会使得重复视频的检测结果相对来说更加准确。

1.2 视频相似性检测流程

目前，国内外已经开发出众多较实用的相似性视频检测系统，如：第1个基于内容的商业化开发检测系统——QBIC（Query by Image Content）系统；能够成功实现基于运动的内容描述系统——VideoQ系统；基于内容检索的视频数字图书馆子项目——Informationdia Digital Video Library项目；面向WWW基于内容的视频检索系统——Web-scope-CBVR系统等[6]。

一个完整的重复视频检测系统应具备如图1所示的几个模块[7]。

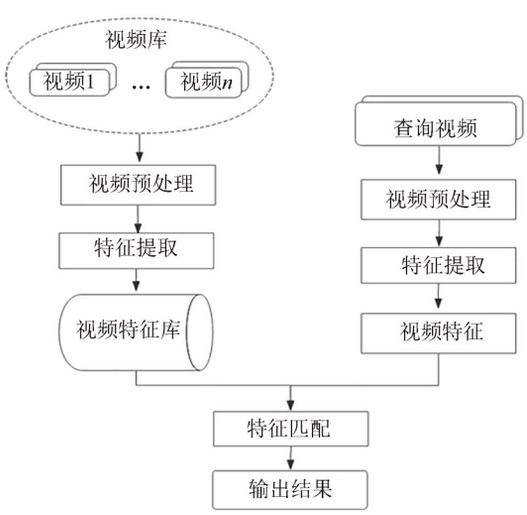


图1 重复视频检测系统框架

Fig. 1 Framework of duplicate video detection system

1.2.1 视频预处理过程

直接对视频流进行分析时，可能由于环境及拍摄的因素导致需要处理的视频帧画面存在一定

问题，如亮度变化、大小和方向不一致等。预处理的目的是使视频质量正常化，消除变换效果，以便在后续流程中进行分析。主要包括镜头分割、灰度化、降噪、裁剪、视频格式转换等处理过程。

1.2.2 特征提取

此步骤为整个重复视频检测流程中最复杂、最关键的一步。特征提取的结果直接影响检测系统的准确度，故众多研究者一直将视频检测算法的重点集中在如何提取视频特征这一步骤上。特征提取主要包括关键帧的视觉特征和镜头运动特征的提取[8]。关键帧即能够反映该镜头主要内容的静态图像集合，其视觉特征的提取主要从颜色、轮廓等几个角度来进行。镜头运动特征的提取可通过分析视频目标的运动轨迹、大小变化等来进行。从特征提取范围的角度，可将视频特征划分为全局特征和局部特征。

然而在实际应用中，需要根据实际情况选择全局特征和局部特征。若处理者对视频的整体内容感兴趣，则用全局特征来描述比较合适。但全局特征的劣势是无法分辨出整个视频的局部细节内容，且提取的特征中会存在较多无用的冗余信息。若想对视频中的某个对象进行具体分析，如判断其运动轨迹，则可利用局部特征的方法，对研究对象边缘的点和线展开分析并依次划分局部区域，即用一些稳定出现的点来代替整个视频，减小计算量。

1.2.3 特征匹配

作为重复视频检测过程的最后一步，主要由搜索和相似度匹配2个步骤组成。搜索即在视频库中通过相应的搜索算法找到与查询视频内容相同的候选视频。此步骤的研究重点在于快速高效搜索算法的构建，如广度优先搜索、B+树搜索等。相似度匹配过程即用来评测查询视频与候选视频的内容重复程度。内容重复越多，则相似性度量值越大，反之越小。相似度匹配的一般方法是距离度量算法，如汉明距离、欧式距离、余弦距离等[9]。得到最终度量结果后与设定阈值进行比较，来确定是否为重复视频。

根据对重复数据进行删除操作的执行位置，可将重复数据删除技术分为基于服务器端和基于客户端的重复数据删除。前者是指客户将需要存储的数据上传至云服务器，再由服务器端进行数据的重复检测及删除操作，该方法会占用大量的上传带宽，也会加大服务器端的计算压力；后者是指客户端在上传数据之前对重复数据进行去重，这样可以大大节省网络带宽和服务器的存储空间，是目前广泛使用的重复数据删除方式[10-13]。

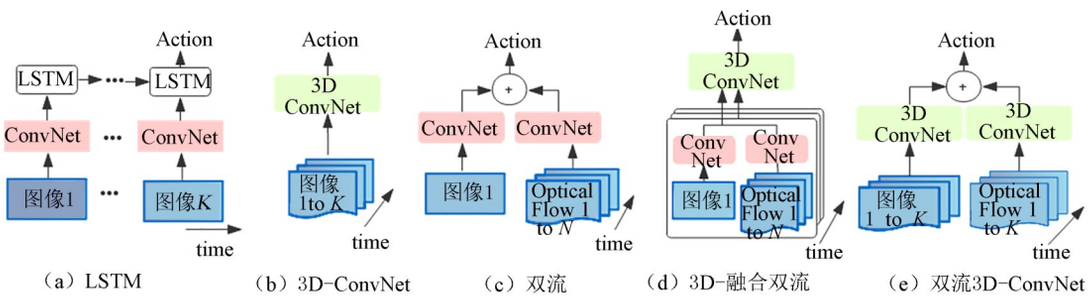
2 基于视频时空信息分类的检测方法

感知哈希（pHash）算法使用离散余弦变换得到视频帧中的低频信息来计算关键帧的指纹及一个64位pHash的二进制序列[10]。作为图片检测和早期视频检测的常用方法，该算法难以应对来自黑客的各类攻击，要么鲁棒性不够，要么执行时间非常高，在指纹提取和匹配操作中通常出发现时间瓶颈。而卷积神经网络近年被广泛应用于计算机视觉中，可以用于图像的分类、检测、分割等，但此过程使用的是二维卷积神经网络。在过去几年中，基于CNN的算法已经取得了比传统方法更好的性能，并且出现了许多出色的CNN模型，例如AlexNet[14]、VGG（Visual Geometry Group）[15]、GoogLeNet[16]等，直接使用CNN模型的全连接层或卷积层的输出作为视频特征。但是，这些CNN特征是在帧级提取的，仍然不够紧凑，而且这些特征没有充分编码视频帧间的重要时间信息，导致性能较差，因此3D CNN就被研究者们提出来了。

当人们将CNN运用到视频分析之初，利用融合卷积神经网络的特征提取方法在ImageNet数据集上的训练效果不理想，人们便怀疑可能是数据量不多导致的。随后谷歌便创建了Sports-1M数据集，包含100多万个视频，但在此数据集上进行训练仍然没有得到较好的结果。这是由于当时人们普遍将视频看作多张图片，进而对每张图片进行2D CNN处理，最后对所有帧的训练结果进行融合，得到最终视频特征。此方法丢失了多张图片之间的时序性运动信息[17-19]，对于情景多变的动态视频来说并不适用。而3D CNN可以更好地描述图片之间的时序性信息。所以，人们对于卷积核层运算符是使用2D内核还是3D内核，网络输入仅仅是视频的RGB形式还是包括光流信息仍存在较大疑虑。因此，便产生了基于视频时空信息分类的相似性视频检测算法，如图2所示[20]。

2.1 ConvNet+LSTM模型

起初，研究者们将视频划分为一帧帧的图像进行特征提取，然后将其合并作为视频特征。此方法虽然思想简单且便于实现，但完全忽略了时



（a）LSTM （b）3D-ConvNet （c）双流 （d）3D-融合双流 （e）双流3D-ConvNet

图 2 视频处理架构模型[20]

Fig. 2 Architecture model of video processing[20]

间结构的问题。从理论上讲，向模型中加入长短期记忆[20]（Long Short-Term Memory，LSTM）循环层，可取得令人满意的结果。可认为LSTM是一种特殊的循环卷积神经网络（Recursive Neural Network，RNN），该结构可以更好地解决长序列在训练过程中梯度消失和梯度爆炸的问题，还可以对状态进行编码，并捕获时间顺序和视频长度的依赖关系。

Carreira等[20]对ConvNet+LSTM模型进行描述，其思想是：将LSTM层进行批量归一化，并放置在Inception-v1的最后一个平均池化层之后，设置隐藏单元为512个。在分类器的顶部添加一个完全连接层，使用所有时间步长上输出的交叉熵损失函数来训练模型，从25帧/s的视频流中，每5帧中保留1帧对输入的视频帧进行下采样。测试期间，只考虑最后一帧的输出。其中，交叉熵常用在分类问题中来衡量真实概率分布与预测概率分布之间的差异程度。一般来说，交叉熵数值越小，说明模型预测效果越好。在分类网络中，通常用softmax激活函数将多个神经元的输出映射到[0, 1]区间内，基于概率论思想使得多个分类预测结果之和为1，再结合交叉熵函数计算损失。

但相比传统的网络而言，由于此网络引入了长期记忆机制，为了避免过拟合，需要更多的参数对模型进行训练，计算量和复杂度较高。而且LSTM网络内部构造不直观，这可能给一些需要掌握网络具体决策过程的金融和医疗领域应用场景带来一定的困扰。

2.2 3D ConvNets模型

3D卷积神经网络具有时空三维滤波器，使用小型3×3×3卷积核的同构体系结构，可直接对时空信息进行提取，卷积层中每个特征图都会与上层中多个临近的连续帧相连接，以捕捉视频的运动信息。但由于其附加内核尺寸的影响，会比2D卷积神经网络具备更多的参数，这也就导致计算过程相对复杂，对于规范化批处理提出了更高的要求，如一个批次可以处理更多的数据。

3D模型最早在文献[21]中被提出并用于行为识别，利用3D特征提取器在视频的时空维度上进行操作，从而捕获视频流中的运动信息，并基于该提取器构建3D CNN模型，组合各种不同3D CNN架构的输出来提高模型的性能。

2.3 双流模型

在卷积层的最后一层加上LSTM结构可以对视频更高层次的变化进行建模，但可能无法捕捉更精细的低层运动，整个训练过程的代价也非常大，因为需要通过多个帧来展开网络进行反向传播，这就必须充分考虑视频帧之间的光流信息。

Facebook基于C3D（Convolutional 3D）网络设计了视频动作分析模型[22]，将网络中的二维滤波器转化为三维滤波器，在时间序列维度上一次性输入连续帧到C3D网络中进行分析。在此网络中，双流法也第一次被提出，该方法将输入的视频流进行2个通道划分，即RGB流和密集光流。双流法将2个卷积神经网络有效融合，得到了有效的分类结果。实验结果表明，3D卷积神经网络是非常好的学习器，对于所有层都使用3×3×3的卷积核效果最好。

后来，Simonyan等[15]提出，首先将单个RGB帧与10个外部计算得到的光流帧的预测值进行平均；然后将其传递给ImageNet预训练网络；最后进行建模。在现有的基准测试中，此类操作可以获得非常高的性能。通俗来讲，该方法将视频训练任务中的卷积神经网络分成两大部分，即空间信息网络和时间信息网络，单独的视频帧集合用于描述视频文件的空间信息，可以理解为互相堆积的图像帧集合，其中包括视频背景、物体等空间特征。而光流信息作为时序信息的载体，输入

到另一神经网络中来提取动作的动态特征，最后将2个模型得到的结果进行融合，作为整个视频的最终特征。

2017年，DeepMind发布视频动作识别模型[20]，将滤波器、内核及各类参数从2D拓展到3D网络中，是一个很深的时空分类器。主要内容包括如下方面。

（1）将2D卷积神经网络拓展至3D。对网络进行三维拓展，这并不仅仅是对时空特性进行重复加工，而是通过膨胀所有的滤波器和池内核来实现的，即赋予2D网络额外的时间维度，将$N\times N$的滤波器变成$N\times N\times N$的。由于视频序列间线性关系的存在，可通过将2D中$N\times N$的滤波器权值在时间维度上重复N次，再除以N，这就在一定程度上保证了卷积滤波器的维度相同。由于图片组成的视频卷积层在时间维度上的输出是恒定的，因此，点状非线性层、平均层和最大池化层均与2D网络结构保持一致。

（2）综合考虑空间、时间和网络深度的增长速度。对于视频流中所有的视频帧，其空间结构（水平与竖直）都是相同的，故其池化内核和步长也应该相同，这就意味着更深层次的特征在空间维度上都受到越来越远的图像位置的同等影响。但由于时间因素的存在，同等条件下都进行接收不符合预期效果。因为会有帧率和图片维度的作用：如果相比空间域，时间域增长太快，可能会将不同目标的边缘合并，影响特征检测过程；如果时间上增长过慢，可能难以捕捉场景动态信息。

（3）Inception-v1体系结构设计。在Inception-v1结构中，第1个卷积层的步长为2，在进行线性分类之前，设置4个步长为2的最大池化层和1个内核为7×7的平均池化层。在实验中，对输入的视频以25帧/s的速度进行处理，在前2个最大池化层中设置内核为1×3×3、步长为1，而在所有其他的最大池化层中使用对称的内核、步长为2，得到了较好的测试结果。最终平均池化层使用2×7×7内核，并使用64帧的片段训练模型，如图3所示，Inception子模块如图4所示。

（4）2个三维网络结构流。虽然3D卷积神经网络可以直接从输入的连续RGB图片中学习到视频运动特征，但这仍然执行的是纯前馈计算，而光流算法在某种意义上是周期性的（如可以对流场进行迭代优化）。故应该以双流结构的思想，一个只接收RGB信息，另一个只接收优化后的平滑光流信息，分别进行训练再对结果取平均。该方案的提出在当时的计算机视觉领域引起了非常大的影响，引导更多学者投入双流3D网络研究中。

2021年，Han等[22]提出基于视频相似性和对齐学习（Video Similarity and Alignment Learning，VSAL）方法，为了减轻空间相似性偏差，将时间相似性建模作为根据帧级空间相似性预测的掩码图。为了进一步定位部分副本，从空间相似性学习步长图，用元素指示时空相似性图上当前部分排列的延伸方向，从遮罩图中获得的起点按照步骤图的指示延伸到部分最优对齐中。通过相似性和对齐学习策略，VSAL在VCDB（VERIS Community Database）核心数据集上获得了最佳的F1分数。此外，通过为FIVR-200k数据集添加新的片段级注释，构建了一个新的部分视频拷贝检测和定位基准，VSAL也验证了其在更具挑战性情况下的有效性。

除了利用卷积内核及视频时空信息分类的方法对相似性视频进行检测之外，也可根据视频的特征来分类。在得到视频特征的基础上，可结合距离度量算法对视频特征的差异性进行刻画。

3 基于视频特征分类的检测方法

从视频特征的角度来看，重复视频数据检测的CNN方法可大致分为2类，即使用视频帧局部

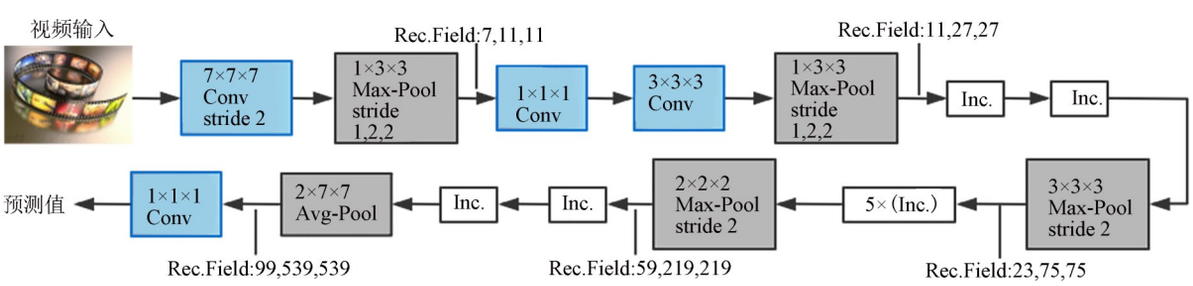


图3 Inflated Inception-v1体系结构[20]

Fig. 3 Architecture of Inflated Inception-v1[20]

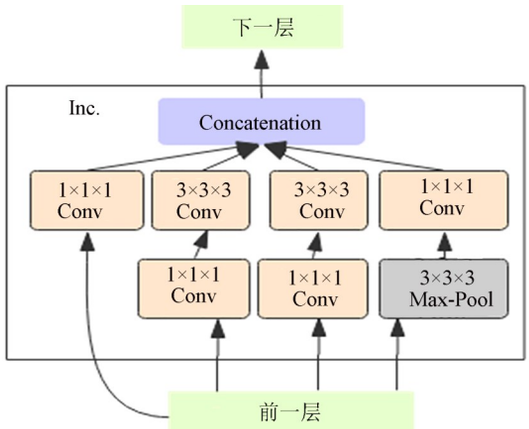


图 4 Inception 子模块[20]

Fig. 4 Submodules of Inception[20]

特征和全局视频特征来计算相似性的方法。之前，人们习惯于使用2D卷积神经网络来逐帧预提取空间特征，可分为帧级/细粒度和视频级/粗粒度视频特征。帧级方法需要精细地将每一对帧进行比较，来计算2个视频之间的相似性，但此方法检索过程通常很慢，因为细粒度机制要求必须存储所有帧的特征，因此需要大量存储空间。而对视频级方法来说，需要进一步对帧级特征进行聚合，以便每个视频仅由1个向量、哈希码或其他形式的索引表示，其组成结构比较紧凑但可能不足以描述视频的细节，所以有时候就需要在帧级和视频级之间做出权衡。

3.1 视频全局特征提取算法

对于使用全局视频特征来计算相似性的方法，通过提取全局视频向量，并使用点积或欧几里得距离来计算视频之间的相似度。Kordopatis等[5]从中间CNN层提取特征创建了可视化代码簿，并基于深度度量学习（Deep Metric Learning，DML）的思想，采用三元丢失方案来训练网络，以学习使相关视频之间的距离最小化并且使不相关视频之间的距离最大化的嵌入方法，并结合汉明距离为整个视频生成哈希码。Song等[23]建立了一个自我监督的视频哈希系统，能够使用编码器-解码器方案捕获帧之间的时间关系。

但根据全局视频特征来计算视频相似性的方案对于需要捕捉更细粒度的视频特征场景来说并不适用，且将完整视频数据参与运算会消耗大量的计算资源，于是出现了局部特征及帧级特征的相似性检测方案。

3.2 视频局部特征提取算法

对于考虑视频帧局部特征的相似性检测方法，通常提取帧级局部特征并进行帧间相似度计算，然后将其聚合为视频级相似度。

Tan等[24]提出通过关键点帧匹配生成基于图的时态网络（Temporal Network，TN）结构，用于检测2个视频之间的最长共享路径。另一种有效方案是利用动态编程（Dynamic Programming，DP）的思想[25]，先计算所有帧对之间的相似度矩阵，继而提取相似度最大的对角线块。为了增加灵活性，还允许有限的水平和垂直运动。Chou等[25]和Liu等[26]将DP和词袋模型（Bag of Words，BOW）相结合来测量帧间相似度。然而由于其刚性的聚合方法，所提出的解决方案不能捕获各种各样的时间相似性模式。

Revaud等[27]提出了循环时间编码（Circulant Temporal Encoding，CTE）的方法，通过傅立叶变换，以时空表示的形式编码帧特征，从而在频域中对视频进行比较。

Poullot等[28]引入了时间匹配内核（Temporal Matching Kernels，TMK），该内核使用帧描述符和时间戳的周期性内核对帧序列进行编码，同时仅参考这些向量来估计时间偏移，这在视频事件检索中取得了当时最佳的性能。然而，由于该方法将视频相似性定义为帧相似性的总和，对噪声的鲁棒性不是很好。因为通常匹配的视频具有连续帧的对应关系，这导致一系列的高相似性，一个简单的帧相似性总和不能区分这种情况和非连续的高相似性，这可能会导致假匹配的结果。

Nie等[29]提出了一种基于层次特征融合的NDVR（Near-duplicate Video Retrieval）视频哈希方法。首先，从视频中提取低级别的手工特征；然后，获得从卷积神经网络中提取的中级深层特征和高级语义特征；最后，将这些语义特征与低级视觉特征相结合，利用层次特征之间发现的全局结构关系和互补性来学习NDVR的哈希码。在CC_WEB_VIDEO数据集上进行大量实验，与谱哈希、多特征融合哈希、随机多视角哈希等技术相比，所提出的框架具有更好的检索性能，其平均检索精度可达0.99。

Wu等[30]认为视频帧质量的退化可能发生在一段较长的时间窗口上，所以忽略时序和局部语义信息，提出特征的融合应当由语义的相似程度引导，而不是时间上的接近程度，来对视频帧进行随机采样并融合。他们将视频看作无序的帧集合，随机采样一定数量的帧，利用某一目标与其他目

标特征的余弦相似度作为权重来引导特征聚合，得到该目标的特征，这样就可以让网络充分学习到每个类别中稳定的特征表达。实验时将SELSA（Sequence Level Semantics Aggregation）模块应用于Faster RCNN的开头部分，训练过程中除了当前帧，还随机从同一视频序列中抽取2帧作为辅助。结果表明，这种方法可以对快速运动的物体带来检测性能上的提升。但由于该方式完全弃用了时序信息，仅利用时间维度上全局的语义信息，这是不全面的。

于是You等[31]使用改进的结构相似性（Structural Similarity，SSIM）和特征图来改进相似度算法，并提出一种SSELSA模块，先通过提出的相似性算法从全局角度选择一些近似帧，再对这些框架的特征进行聚集。该方法可以更全面地比较相似性，并在不增加过多冗余计算的情况下降低聚类错误的风险，还利用特征金字塔和动态区域感知卷积（DR-Conv）来增强Faster RCNN的特征提取能力。

在视频检索领域，为获得有效的视频表示，大多数视频检索系统需要大量手动注释的数据进行训练，这使得处理代价很高，效率低下。此外，大多数检索系统都是基于帧级特征进行视频相似性搜索的，会造成极大的存储和搜索压力。

为了解决上述问题，He等[32]提出一种视频表示学习（Video Representation Learning，VRL）方法。先通过对比学习从未标记的视频中有效地学习视频特征，以避免手动注释的昂贵代价。再利用transformer结构将帧级特征聚合到剪辑级，以减少存储空间和搜索复杂性。这样可以从剪辑帧之间的交互中学习互补和判别信息，并获得帧排列和缺失不变能力，以支持更灵活的检索方式。在具有挑战性的近重复视频检索数据集FIVR-200 k和SVD（Short Video Dataset）上进行验证，该方法在准确性和效率上可实现视频检索的最佳性能。

2023年，Tan等[33]提出一种快速的视频拷贝检测框架。在这个框架中，参考视频的所有帧都保存在一个K近邻分类（K-Nearest Neighbor，KNN）可搜索数据库中，当需要对某个视频片段进行查询时，不需要扫描所有参考视频，只需在KNN数据库中进行快速检索，生成一个包含所有候选视频的名单，依此定位拷贝片段。在VCDB数据集上对该算法进行评估，其F1基准分数大大超过当时最先进的水平，算法速度也得到了显著提高。

4 基于视频多模态的深度学习方法

2019年，Kordopatis等[34]提出了一种视频相似性学习架构Visil，综合考虑了视频对之间的细粒度时空关系，可以从细化的帧到帧相似性矩阵中计算视频到视频的相似性。并将所有视频帧之间的相似性矩阵馈送到4层CNN中，使用倒角相似性（Chamfer Similarity，CS）将其计算为视频到视频的相似性得分，这可以有效避免在视频之间的相似性计算之前进行特征聚合。该方案使用三重损失的训练方法，在5个公共基准数据集上进行评估表明，与现有技术相比，检测性能得到了较大提升。

但Liu等[35]发现，三重损失方案存在一些缺陷，如：（1）划分锚点视频、正视频、负视频的分批策略虽然简单方便，但可能会出现采样偏差，因为批次内的负样本可能具有许多语义上相似的视频和文本内容，会混淆模型的判断；（2）负样本的数量会受到批处理大小设置的限制，对比学习的性能越好，需要设置的批处理数量就越大，然而，批处理大小的设置取决于硬件平台提供的内存大小；（3）与单编码器解决方案相比，由于模态差距和双编码器架构带来的大量参数，就会导致大量的计算成本和模型训练的收敛障碍。

针对以上问题，Liu等[35]提出了一种基于统一Transformer的去偏动量对比学习框架（Debiased Mormentum Contrastive Learning，DMCL）。使用视频文本对之间的上下文语义相似性作为软标签监督，以减轻假阴性带来的不利影响，利用基于Transformer的统一编码器VideoSim来学习多模态表示。通过4个子任务进行预训练，包括视频标签分类、掩码语言建模、掩码帧建模和帧文本匹配。在MVSE数据集[36]上进行了比较实验，以衡量短视频的多模态相似性。结果表明，将视频文中标题、标签等信息的特征进行融合所得到的测试性能高于单模态。

目前，多模态学习方法已逐步应用到视频目标识别、行为预测、相似度检测等领域。对于多模态学习方法，其多模态数据值的有效性取决于提取方案的设计。在特定应用场景中，为了更高效地完成任务，模式数据的选取也非常关键。广

泛的研究证明，有效的多模态融合方法有助于突出单个数据模式各自的信息作用，同时巩固其任务优势。现有的大多数多模态深度学习方法依赖于注意力的多模态融合策略[30]、Transformer融合策略[37]、图的多模态融合策略[32-33]等。

Bahdanau等[38]在2015年首次提出注意力机制，可以使网络模型的解码器在每次解码的过程中，能够针对性地选择源语言中“对齐”的词来指导目标语言的解码，包括全局注意力和局部注意力2种方法。2017年，Vaswani等[39]提出了由多头注意力和自注意力等模块构成的Transformer，目前已广泛应用在计算机视觉、语言识别等领域。2018年，Anderson等[40]提出一种“自底向上”和“自顶向下”相结合的注意力机制，此方法能够让注意力在对象和其他突出区域级上更自然地进行计算。自顶向下的注意力模型可以提取图像特征，类似于对图像进行特征编码，实现编码阶段任务；而自底向上的注意力模型用于学习调整特征权重，时刻关注图像内容，相当于解码阶段。但这种方式是基于文本表示选择性地关注图像中的兴趣区域。Lu等[41]认为文本注意力和视觉注意力同等重要，因此提出了协同注意力机制，用图像表征来引导问题注意，用问题表征来引导图像注意，使得模型在图像和问题之间具有自然的对称性。后来，受Transformer模型的启发，Yu等[42]提出了一个深度模块化共同注意网络（Modular Co-attention Networks，MCAN），是由深度级联的模块化共同注意层（MCA）组成的。每个MCA层利用2个基本注意单元的模块化组合，可实现文本中的任一词与图像中的任一区域间的完全交互。

Transformer具有非常强大的特征学习能力，2019年被应用到多模态领域。基于Transformer的多模态融合算法又分为单流模型[37, 43]和双流模型[44]。单流模型使用一个Transformer在刚开始时便对多模态信息进行充分融合；而双流模型则将Transformer编码应用在不同的模态上，再通过协同注意力机制来实现不同模态特征的融合，可以使用不同模态独立地处理需求。但目前没有更多的对比实验来分析单流模型和双流模型的优缺点。为了进一步提升模型的检测性能，研究者们将CNN在局部特征提取方面的优势与Transformer在全局信息建模方面的优势相结合，提出了CNN-Transformer混合架构，此研究方向已成为当下计算机视觉领域研究中的热点主题。如Yuan等[45]提出了CeiT（Convolution-enhanced Image Transformer）架构，将CNN有效融入视觉Transformer架构中，以提取低层次特征、加强局部性和建立长范围的依赖关系。实验结果表明，CeiT具有更好的效果和泛化能力，无需大量训练数据和额外的CNN teacher模型。此外，CeiT模型还表现出了更好的收敛性，可以显著降低训练成本。

相比于CNN、RNN等神经网络模型，图神经网络可以用来处理更加复杂的数据，更加精准和灵活地对现实应用中的数据进行建模。如Yin等[46]提出将基于图的多模态融合编码器应用到多模态神经网络翻译模型中，把图和句子用一个统一多模态图来表示，再基于该图堆叠生成多模态融合层，这些层之间进行反复迭代，来表示节点之间的语义交互，生成图编码。该方法可以同时学习模态内和模态间的各种颗粒度语义关系，进而显著提高机器翻译的性能。除此之外，基于图的多模态融合模型也逐渐被应用到情感识别。如Hu等[47]提出MMGCN（Multimodal Fusion Via Deep Graph Convolution Network）模型来对对话中的上下文信息、多模态信息、对话者信息和长距离信息进行建模，可以通过说话人向量把其音色特征融入到情感识别模型中。就图融合模型而言，如何设计更加合理的图神经网络模型，利用图神经网络在信息传播和推理上的优势，改善文本挖掘、推荐系统和计算机视觉等领域的应用效果，是人工智能领域研究的一个热点问题。

多模态检测算法在视频相似度领域的应用相对较少，较为经典的算法如2021年AI算法大赛中的各类视频相似性检测算法。其中，排名第一的方案[48]原理是通过训练得出的Embedding模型来计算视频之间的余弦相似度，用斯皮尔曼秩相关系数来进行评估，并微调模型。最终，通过这种方式，测试集上最好单模的斯皮尔曼秩相关系数为0.836，集成2个模型时达到0.845，3个模型时达到0.849，5个模型时达到0.852。

5 数据集与性能评估

5.1 数据集

视频相似性检测领域的代表性数据集信息如表1所示。

TRECVID[49]数据集包含200 h时长的电视节目视频，约2 000个查询片段。查询片段是对原始

表 1 常见数据集信息统计表

Tab. 1 Information of common datasets

数据集	视频数/个	分类数	发布年份	特点
CC_WEB_VIDEO[50]	12 790	24	2007	均来源于真实网络，包含24个主题
UQ-VIDEO[54]	132 647	—	2011	是对CC_WEB_VIDEO数据集的拓展，规模大，但查询视频的数量有限
HMDB51[55]	6 849	51	2011	规模小，方便使用，包含51类人体动作
TRECVID[49]	22 759	—	2011	视频类型丰富、数量较多，如新闻杂志、纪录片、档案视频、电影等
UCF101[56]	13 320	101	2012	包含101个人类行为类别
VCDB[53]	100 528	—	2014	规模小，未经过人为的变化处理，重复关系已进行标注
YouTube-8M[57]	8 264 650	4 800	2016	规模大、类别广，以长视频为主
SVD[51]	超过530 000	—	2019	规模大，设计了视频的时间和空间变换，视频变体丰富
VCSL[52]	超过440 000	—	2022	类别广、视频的时间范围较为精确

视频进行复制、再编码等操作得来的，涉及的领域包括文化、新闻、纪录片、教育等方面。该数据集利用11 256个查询视频构建查询集，对查询视频进行一些预定义转换，生成11 503个标记视频，因此，该数据集共收集了22 759个视频。CC_WEB_VIDEO[50]是视频重复检测领域公开的四大数据集之一。该数据集包含了24个不同的关键词，统计了谷歌、YouTube和雅虎三大视频网站上的部分文件，总计达12 790个视频，涵盖音乐、电影、体育、动画等多种类型，其相似视频主要涉及帧率、格式、分辨率、颜色以及添加字幕等变化。另外，该数据集最大的特点是所有视频均来自于真实网络，没有对其进行模拟拷贝的转化操作，可认为该数据集能够体现真实网络的拷贝情况。

此外，由于收集和标记近乎重复的长视频的成本高昂，因此，大多数视频集都是小规模的，缺乏多样性。2019年，Jiang等[51]构建了一个专门用于相似性视频检测任务的大型短视频数据集SVD，包含超过500 000个短视频和超过30 000个近乎重复的标签视频，他们使用多种视频挖掘技术来构建正/负视频对。此外，还设计了时间和空间变换，以模拟实际应用程序中的用户攻击行为，从而构建更复杂的SVD变体。

He等[52]提出了一个新的视频标注数据集VCSL（Video Copy Segment Localization），不仅包含16万个真实的视频拷贝数据，还有超过28万个局部拷贝的片段对，涵盖的视频类别广，视频持续时间范围较为广泛，每个收集到的视频都附有精确的开始和结束时间戳，非常适用于拷贝视频片段定位的算法要求。

VCDB[53]数据集是上海智能信息实验室和复旦大学收集的小规模视频相似性检测数据集，包括528个核心视频、9 236个复制片段的核心视频以及10万多个视频片段的未标注数据集，覆盖商业、电影、运动等多个主题，这些视频没有经过人为的变化处理，重复关系已进行标注，广泛应用在计算机视觉、目标检测等研究领域。

UQ_Video数据集[54]是对CC_WEB_VIDEO数据集的拓展。作者利用24个查询视频和12 790个标记视频作为UQ_Video的查询集和标注集，构建了一个包含119 833个视频的背景干扰集，最终共收集了132 647个视频。虽然该数据集的规模比CC_WEB_VIDEO更大，但查询数量有限。所以，对于大部分检测算法，其性能相对较弱。

HMDB51数据集[55]是一个非常小型、方便使用的数据集，来源于YouTube、Google等网站，共包含51类动作，共计6 849个视频，约2 GB。其类别主要包括面部动作、面部操作、人体动作、身体动作等。

UCF101数据集[56]来源于YouTube网站，由于包含101种不同的视频类别，故而称为UCF101。其中包含13 320个短视频、5大类动作，分别为人与物体交互、肢体动作、人与人交互、音乐器材、各类运动的场景。

5.2 评价标准

全类平均正确率P_{mA}是目标检测和计算机视觉等领域十分重要的衡量指标，来分析不同模型或方法的性能，是通过对各个类别检测的平均正确率P_{A}进行综合加权平均得到的，而P_{A}是基于召回率和精确率来计算的。

召回率（P_{recall}）是指实际重复的视频被判定

为重复视频的数量占全部重复视频的比例，反映了视频重复数据删除系统检索的正确性。对于重复数据删除系统而言，召回率越高意味着系统重复检测的准确性越高。

精准率（$P_{\text{precision}}$）表示测试结果中判定为重复的视频占总的实际重复视频的比例。

如果总的视频数量为n，则平均正确率定义为

$$P_{\text{A}}=\sum_{i=1}^{n}P_{\text{precision},i}\left(P_{\text{recall},i+1}-P_{\text{recall},i}\right) \tag{1}$$

则P_{mA}的值为

$$P_{\text{mA}}=\frac{1}{Q}\sum_{q=1}^{Q}P_{\text{A},q} \tag{2}$$

式中：Q为总的类别数。

除了上述评价指标外，误报率（P_{FR}）也经常被作为一种有效的性能评价指标，是指实际非重复视频被错误检测为重复的数量占全部非重复视频的比例，这反映了视频重复数据去重过程中的可靠性。对于重复数据删除系统而言，误报率越低，系统可靠性越强。

5.3 实验评估

近年来出现的视频相似性检测方法以卷积神经网络为主。表2主要针对典型的检测算法性能进行比较分析，所有精度数据均来自于相应文献。

表2 不同视频相似性检测方法性能比较

Tab. 2 Performance comparison of different video similarity detection methods

检测方法	数据集	P_{mA}/%	备注
DH[59]	UCF101	47	Disaggregation Hashing(分解哈希)，是一种简单高效的高维视频数据哈希方案
SRH[60]	UCF101	75.4	Supervised Recurrent Hashing(监督递归哈希)，是通过深度神经网络获得的特征来设计的哈希方案
DVH[61]	UCF101	71.2	Deep Video Hashing(深度视频哈希)，是一种用于可扩展视频搜索的深度视频哈希方法
CSQ[62]	UCF101	87.4	Central Similarity Quantization(中心相似性量化)，它优化了数据点与其哈希中心之间的中心相似性
DH[59]	HMDB51	31	—
SRH[60]	HMDB51	50.9	—
DVH[61]	HMDB51	51.8	—
CSQ[62]	HMDB51	57.9	—
ViSil[34]	CC_WEB_VIDEO	99.3	不支持时间对齐
TN[24]	CC_WEB_VIDEO	98.7	支持时间对齐
DP[25]	CC_WEB_VIDEO	98.2	支持时间对齐
SPD[63]	CC_WEB_VIDEO	98.5	Similarity Pattern Detection(相似性检测模块)，该模块具有鲁棒，可用于时间对齐，支持时间对齐
DML[50]	CC_WEB_VIDEO	97.9	支持时间对齐
TCA-C[64]	CC_WEB_VIDEO	98.3	Temporal Context Aggregation-Cosine(时间上下文编码-余弦距离)，支持时间对齐
PPT[25]	CC_WEB_VIDEO	95.8	Pattern-Based Indexing Tree(PI-tree)+M-Pattern-Based Dynamic Programming (mPDP)+Time-Shift M-Pattern Similarity(TPS) 基于模式的索引树(PI-tree)+基于m-模式的动态规划(mPDP)+时移m-模式相似性(TPS)
MFH[54]	CC_WEB_VIDEO	95.4	Multiple Feature Hashing(多特征哈希)
LBP[65]	CC_WEB_VIDEO	95.3	Local Binary Pattern(局部二进制模式)
HIRACH[50]	CC_WEB_VIDEO	95.2	Hierarchical(分层模式)，具有非常高的精确率和召回率
SIG-CH[50]	CC_WEB_VIDEO	89.2	Global Signature on Color Histogram(颜色直方图上的全局标记)，此算法对于具有简单场景或具有较小编辑和变化的复杂场景的查询实现了良好的性能
PI-tree[25]	UQ_Video	79.2	Pattern-Based Indexing Tree(基于模式的索引树) 算法检测时间2.88 s
PPT[25]	UQ_Video	88.3	算法检测时间5.33 s

由表2可以看到，对于UCF101数据集和HMDB51数据集，与DH方案、SRH方案、DVH方案相比，CSQ方案的性能都有显著提升。CSQ方法在UCF101上取得更大改进主要是由于其严重的数据不平衡。而DH方案的表现较差，在2种数据集上，其P_{mA}值都未达到50%。

ViSil方案、TN方案、DP方案及SPD方案的视频相似性检测性能在公开数据集CC_WEB_VIDEO上的效果都非常好，且各方案间的P_{mA}值相差不大。但具有最高P_{mA}值的ViSil不能输出准确的时间位置，而其他3种方法具有对齐能力，这也是ViSil方案的缺陷之一。

在CC_WEB_VIDEO数据集上，PPT方案的效果比ViSil等方案较弱，但比MFH方案、LBP方案、HIRACH方案和SIG_CH方案优秀，而SIG_CH方案是由于直接采用颜色直方图而导致性能不佳。

在UQ_Video数据集上，PI-tree和PPT这2种方案的有效性比CC_WEB_VIDEO数据集稍微逊色，主要是因为UQ_Video数据集中有10多万个干扰视频。其中，PPT方案比PI-tree方案性能好，但时间开销较大。

6 未来发展趋势

目前，视频相似性检测技术已经取得了非常大的突破，但仍有众多值得深入研究的问题。

（1）较低的检测成本。由于现今大多数检测算法（如多模态数据分析等）都是基于神经网络的，而利用此方式去处理占用存储空间非常高的视频数据会引入很大的计算开销，尤其是网络层数特别深时。如何有效减小模型的训练及测试开销非常重要，通用做法是从时间维度入手。一方面，相邻帧之间往往具有较大的相似性，逐帧处理将引入众多冗余计算；另一方面，并非每个视频帧都是和检测任务相关的，因此，动态地对视频关键帧进行提取并进行重点处理，可以降低计算成本。此外，也可以从空间维度来分析，因为对于每个视频帧而言，并不是所有的区域都和算法研究相关。例如目标检测，可以通过对重点目标所在区域进行处理，提取其局部特征，再进行下一步计算。因此，综合考虑多模态视频特征时的计算复杂度问题很值得进行深入思考。

除此之外，众多云服务提供商也推出了云计算资源，开发者可以适当选择虚拟云服务产品，来减少本地硬件部署的经济成本和时间成本等。由于轻量级服务器及各类中间件等拥有高可用性、高性能，使用方便、稳定、可靠，提供了多种配置选项，可以有效避免各类环境包引起的冲突，帮助用户快速构建算法，使得越来越多的人员参与到模型研发中去，而不仅仅局限于相关科研人员，这将是未来算法部署的发展趋势。

（2）自适应的检测阈值。现有的很多重复视频检测方案所采用的比较阈值都是基于经验得到的经验阈值，若在机器学习的背景下，能够自适应生成符合各类检测标准的阈值，将会给算法间的比较带来极大帮助。

（3）完整的检测体系。视频检测技术相关的配套体系建设还有一定的发展空间。例如：怎样去设计安全的视频相似性检测算法，能够有效抵抗黑客的攻击？云服务器如何识别并保存客户端上传的高质量视频文件，即视频的质量如何评估？检测任务完成后，对于相似的和不相似的视频又该如何处理？是否还有继续研究的价值？这些方面仍需进行更加深入的研究。

（4）通用的检测大模型。虽然大部分相似性检测算法能够取得比较完美的效果，但从表2可以看到，对于不同数据集而言，其检测性能差异较大，因此，研究一种通用的、可以针对多种不同数据集的高效大模型至关重要。目前，已经有很多专家提出了多模态视频检测方法，但在检测性能和效率方面仍存在较大的矛盾，因此，在各模态特征的提取及特征融合等方面仍存在很大的改进空间。研究改进或替代Transformer的高效模型、超大规模模型分布式并行训练、多模态信息之间更细粒度的对齐建模、多模态感知下决策一体化的新一代机器人技术等，都是未来该领域的热点问题。除此之外，如果还需要更加精确地对视频重复片段进行精确定位，如精确到毫秒级等，就要研究更加细粒度的拷贝片段定位算法。

参考文献：

[1] LI D，ZHU B Q，XU K L，et al. Enhanced cross-modal transformer model for video semantic similarity measurement[J]. IEEE Transactions on Circuits and Systems，II：Express Briefs，2024，71（1）：475-479.

[2] LI G B，XIE Y，LIN L，et al. Instance-level salient object segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ，USA：IEEE，2017：247-256.

[3] ZHU H D，WEI H R，LI B Q，et al. A review of video object detection：Datasets，metrics and methods[J]. Applied Sciences，2020，10(21)：7834.

[4] JIANG Y G，WANG J J. Partial copy detection in videos：A benchmark and an evaluation of popular methods[J]. IEEE Transactions on Big Data，2016，2(1)：32-42.

[5] KORDOPATIS-ZILOS G，PAPADOPOULOS S，PATRAS I，et al. Near-duplicate video retrieval by aggregating intermediate CNN layers[C]//International Conference on Multimedia Modeling. Cham：Springer，2017：251-263.

[6] RASHID F，MIRI A，WOUNGANG I. Proof of storage for video deduplication in the cloud[C]//2015 IEEE International Congress on Big Data. Piscataway，NJ，USA：IEEE，2015：499-505.

[7] RAZAVIAN A S，SULLIVAN J，CARLSSON S，et al. Visual instance retrieval with deep convolutional networks[J]. ITE Transactions on Media Technology and Applications. 2016，4(3)：251-258.

[8] CHEN J J，LI S，LIU D H，et al. Indoor camera pose estimation via style-transfer 3D models[J]. Computer-Aided Civil and Infrastructure Engineering，2022，37(3)：335-353.

[9] 刘妍. 基于Lucene的余弦距离检测文档相似度方法的研究[J]. 信息系统工程，2014(4)：129-130，142.
LIU Y. Research on cosine distance detection document similarity method based on Lucene[J]. China CIO News，2014(4)：129-130，142.

[10] GUZMAN-ZAVALETA Z J，FEREGRINO-URIBE C. Towards a video passive content fingerprinting method for partial-copy detection robust against non-simulated attacks[J]. PLoS ONE，2016，11(11)：e0166047.

[11] DAWOOD Q M，RAO K G，RAO B B，et al. Secure video data deduplication in the cloud storage using compressive sensing[J]. International Journal of Computer Science Trends and Technology，2018，6(3)：38-46.

[12] YELA D F，STOWELL D，SANDLER M. Spectral visibility graphs：Application to similarity of harmonic signals[C]//Proceedings of 27th European Signal Processing Conference. Piscataway，NJ，USA：IEEE，2019：1-5.

[13] WANG Z B，ZHU Y M，ZHANG Q M，et al. Graph-enhanced spatial-temporal network for next POI recommendation[J]. ACM Transactions on Knowledge Discovery from Data，2022，16(6)：1-21.

[14] KRIZHEVSKY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM，2017，60(6)：84-90.

[15] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J/OL]. Computer Science，2014[2023-12-08]. http：//arxiv.org/abs/1409.1556.pdf.

[16] SZEGEDY C，LIU W，JIA Y Q，et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway，NJ，USA：IEEE，2015：1-9.

[17] YANG W J，WANG W M，WANG Q Y，et al. Image retrieval method based on perceptual hash algorithm and bag of visual words[J]. Journal of Graphics，2019，40(3)：519-524.

[18] TAN H K，NGO C W，CHUA T S. Efficient mining of multiple partial near-duplicate alignments by temporal network[J]. IEEE Transactions on Circuits and Systems for Video Technology，2010，20(11)：1486-1498.

[19] HE K M，ZHANG X Y，REN S Q，et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(9)：1904-1916.

[20] CARREIRA J，ZISSERMAN A. Qua vadis，action recognition? A new model and the kinetics dataset[EB/OL]. 2017[2023-12-08]. http：//arxiv.org/abs/1705.07750.

[21] JI S W，XU W，YANG M，et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(1)：221-231.

[22] HAN Z，HE X T，TANG M Q，et al. Video similarity and alignment learning on partial video copy detection[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York：ACM，2021：4165-4173.

[23] SONG J K，ZHANG H W，LI X P，et al. Self-

supervised video hashing with hierarchical binary auto-encoder[J]. IEEE Transactions on Image Processing, 2018, 27(7): 3210-3221.

[24] TAN H K, NGO C W, HONG R, et al. Scalable detection of partial near-duplicate videos by visual-temporal consistency[C]//Proceedings of the 17th ACM International Conference on Multimedia. New York: ACM, 2009: 145-154.

[25] CHOU C L, CHEN H T, LEE S Y. Pattern-based near-duplicate video retrieval and localization on web-scale videos[J]. IEEE Transactions on Multimedia, 2015, 17(3): 382-395.

[26] LIU H, ZHAO Q J, WANG H, et al. An image-based near-duplicate video retrieval and localization using improved Edit distance[J]. Multimedia Tools and Applications, 2017, 76(22): 24435-24456.

[27] REVAUD J, DOUZE M, SCHMID C, et al. Event retrieval in large video collections with circulant temporal encoding[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2013: 2459-2466.

[28] POULLOT S, TSUKATANI S, NGUYEN A P, et al. Temporal matching kernel with explicit feature maps[C]//Proceeding of the 23rd ACM International Conference on Multimedia. New York: ACM, 2015: 381-390.

[29] NIE X S, LIN P G, YANG M Z, et al. Hierarchical feature fusion hashing for near-duplicate video retrieval[[J]. Scientia Sinica Informationis, 2018, 48(12): 1679-1708.

[30] WU H P, CHEN Y T, WANG N Y, et al. Sequence level semantics aggregation for video object detection[C]//2019 IEEE / CVF International Conference on Computer Vision (ICCV). Piscatavway, NJ, USA: IEEE, 2019: 9216-9224.

[31] YOU H T, LU Y F, TANG H H. Improved feature extraction and similarity algorithm for video object detection[J]. Information, 2023, 14(2): 115.

[32] HE X T, PAN Y L, TANG M Q, et al. Learn from unlabeled videos for near-duplicate video retrieval[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2022: 1002-1011.

[33] TAN W J, GUO H W, LIU R S. A fast partial video copy detection using KNN and global feature database[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2022: 459-467.

[34] KORDOPATIS-ZILOS G, PAPADOPOULOS S, PATRAS I, et al. ViSiL: Fine-grained spatio-temporal video similarity learning[C]//2019 IEEE / CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 6350-6359.

[35] LIU K H, WANG J, ZHANG X J. Debiased momentum contrastive learning for multimodal video similarity measures[J]. Neurocomputing, 2024, 563: 126938.

[36] ZENG Z Y, LUO Y S, LIU Z H, et al. Tencent-MVSE: A large-scale benchmark dataset for multi-modal video similarity evaluation[C]//2022 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 3128-3137.

[37] SU W J, ZHU X Z, CAO Y, et al. VL-BERT: Pretraining of generic visual-linguistic representations [EB / OL]. 2019[2023-12-08]. http://arxiv.org/abs/1908.08530.

[38] BAHDANAU D, CHO K H, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//Proceedings of the 3rd International Conference on Learning Representations. New York: ICLR, 2015.

[39] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5998-6008.

[40] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//2018 IEEE / CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 6077-6086.

[41] LU J S, YANG J W, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 289-297.

[42] YU Z, YU J, CUI Y H, et al. Deep modular co-attention networks for visual question answering [C]//2019 IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Piscataway, NJ, USA: IEEE, 2019: 6274-6283.

[43] LI X J，YIN X，LI C Y，et al. Oscar：Object-semantics aligned pre-training for vision-language tasks[C]//Proceedings of the European Conference on Computer Vision. Cham：Springer，2020：121-137.

[44] YU F，TANG J J，YIN W C，et al. ERNIE-ViL：Knowledge enhanced vision-language representations through scene graphs[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(4)：3208-3216.

[45] YUAN K，GUO S P，LIU Z W，et al. Incorporating convolution designs into visual transformers[C]//2021 IEEE/CVF International Conference on Computer Vision（ICCV）. Piscataway，NJ，USA：IEEE，2021：559-568.

[46] YIN Y J，MENG F D，SU J S，et al. A novel graph-based multi-modal fusion encoder for neural machine translation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg，PA，USA：Association for Computational Linguistics，2020：3025-3035.

[47] HU J W，LIU Y C，ZHAO J M，et al. MMGCN：Multimodal fusion via deep graph convolution network for emotion recognition in conversation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing（Volume 1：Long Papers）. Stroudsburg，PA，USA：Association for Computational Linguistics，2021：5666-5675.

[48] MA Z R，LOU M J，OUYANG X. Top1 solution of QQ browser 2021 AI algorithm competition track 1：Multimodal video similarity[EB/OL]. 2021[2023-12-08]. http：//arxiv.org/abs/2111.01677.

[49] 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，2011：10-12.

[50] WU X，HAUPTMANN A G，NGO C W. Practical elimination of near-duplicates from web video search[C]//Proceedings of the 15th ACM International Conference on Multimedia. New York：ACM，2007：218-227.

[51] JIANG Q Y，HE Y，LI G，et al. SVD：A large-scale short video dataset for near-duplicate video retrieval[C]//2019 IEEE/CVF International Conference on Computer Vision（ICCV）. Piscataway，NJ，USA：IEEE，2019：5280-5288.

[52] HE S F，YANG X D，JIANG C，et al. A large-scale comprehensive dataset and copy-overlap aware evaluation protocol for segment-level video copy detection[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，NJ，USA：IEEE，2022：21054-21063.

[53] JIANG Y G，JIANG Y D，WANG J J. VCDB：A large-scale database for partial copy detection in videos[C]//Proceedings of the European Conference on Computer Vision. Berlin：Springer，2014：357-371.

[54] SONG J K，YANG Y，HUANG Z，et al. Multiple feature hashing for real-time large scale near-duplicate video retrieval[C]//Proceedings of the 19th ACM international conference on Multimedia. New York：ACM，2011：423-432.

[55] KUEHNE H，JHUANG H，GARROTE E，et al. HMDB：A large video database for human motion recognition[C]//2011 International Conference on Computer Vision. Piscataway，NJ，USA：IEEE，2011：2556-2563.

[56] SOOMRO K，ZAMIR A R，SHAH M. UCF101：A dataset of 101 human actions classes from videos in the wild[J/OL]. Computer Science，2012[2023-12-08]. http：//arxiv.org/abs/1212.0402.

[57] ABU-EL-HAIJA S，KOTHARI N，LEE J，et al. YouTube-8M：A large-scale video classification benchmark[EB/OL]. 2016[2023-12-08]. http：//arxiv.org/abs/1609.08675.

[58] 刘先梁，徐建，郝沛健，等. 基于改进Faster R-CNN的绝缘子及其缺陷检测[J]. 信息与电脑，2023，35(6)：78-81.
LIU X L，XU J，HAO P J，et al. Detection of insulators and its defects based on improved Faster RCNN[J]. Information & Computers. 2023，35(6)：78-81.

[59] QIN J，LIU L，YU M Y，et al. Fast action retrieval from videos via feature disaggregation[J]. Computer Vision and Image Understanding，2017，156(C)：104-116.

[60] GU Y，MA C，YANG J. Supervised recurrent hashing for large scale video retrieval[C]//Proceedings of the 24th ACM international conference on Multimedia. New York：ACM，2016：272-276.

[61] LIONG V E，LU J W，TAN Y P，et al. Deep video hashing[J]IEEE Transactions on Multimedia. IEEE，2017，19(6)：1209-1219.

[62] YUAN L，WANG T，ZHANG X P，et al. Central similarity quantization for efficient image and video retrieval[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Piscataway，NJ，USA：IEEE，2020： 3080-3089.

[63] JIANG C，HUANG K M，HE S F，et al. Learning segment similarity and alignment in large-escale content-based video retrieval：Computer Vision and Pattern Recognition[C/OL]//2023 IEEE/CNF Conference on Computer Vision and Pattern Recognition（CVPR）. [2023-10-20]. http：//doi.org.10.48550/arXiv.2309.11091.

[64] SHAO J，WEN X，ZHAO B C，et al. Temporal context aggregation for video retrieval with contrastive learning[C]//2021 IEEE/CVF Winter Conference on Applications of Computer Vision（WACV）. Piscataway，NJ，USA：IEEE，2020：3267-3277.

[65] SHANG L F，YANG L J，WANG F，et al. Real-time large scale near-duplicate web video retrieval[C]//Proceedings of the 18th ACM International Conference on Multimedia. New York：ACM， 2010：531-540.

通信作者及其研究团队介绍

王忠，教授、博士生导师。长期从事计算机学科教学科研工作，获国防科技进步二等奖1项，军队科技进步一等奖1项、二等奖4项，出版教材5部，发表学术论文50余篇。目前主要从事高性能嵌入式智能计算、边缘智能以及智能软件可靠性研究工作。

本科研团队主要由火箭军工程大学基础部相关教研室人员构成，针对新一代航空航天嵌入式智能系统对算力、存储以及信息交换能力的需求以及物理约束和使用条件约束，开展先进智能系统体系结构、智能边缘计算以及嵌入式智能系统可靠性与可信性设计与验证评估技术研究，主要面向两个领域开展科研工作。

1.先进元器件与智能微系统理论与集成技术。针对航空航天装备电子系统智能化发展以及平台嵌入式环境对实时性、功耗、体积等方面的约束，开展基于自主可控的嵌入式智能系统设计技术研究，包括数模混合电路设计、嵌入式智能芯片与核心器件、嵌入式神经网络、嵌入式光计算与光互联、嵌入式支撑软件等，聚焦基于国产芯片的电路系统硬软件设计、嵌入式智能计算芯片体系结构、深度神经网络模型选择压缩与优化、智能计算加速平台架构及嵌入式AI软硬件协同设计关键技术，为解决资源受限的嵌入式环境下深度神经网络的部署与优化问题、人工智能应用小型化等问题探索可行的技术途径。

2.智能边缘计算技术。针对特定应用场景对边缘节点计算能力、存储能力、学习和推理能力的需求，开展基于AIoT的智能边缘计算技术研究，包括体系结构、AI模型与算法、互联网络等，聚焦面向不同设备的模型压缩和优化、基于异构硬件资源的系统优化、主动持续训练与学习、隐私保护和模型安全等关键技术，为解决深度学习模型在资源受限的边缘设备运行问题、用户隐私数据保护以及模型训练使用的一体化问题等问题探索可行的技术途径。