中国科学: 信息科学 2021 年 第 51 卷 第 5 期: 695–734
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

视频萃取

李学龙*, 赵斌

西北工业大学光电与智能研究院, 西安 710072
* 通信作者. E-mail: li@nwpu.edu.cn

收稿日期: 2020–06–06; 修回日期: 2020–11–06; 接受日期: 2021–02–22; 网络出版日期: 2021–04–13

国家重点研发计划 (批准号: 2018AAA0102201)、国家自然科学基金 (批准号: 61871470, 61761130079, U1801262) 和博士后科学基金 (批准号: 2020TQ0236) 资助项目

摘要 视频数据是人们日常生活中最重要的信息载体之一. 视频萃取 (video distillation) 通过研究视频数据的时空和语义特性, 探索简洁高效的数据展示形式和信息感知模态, 是计算机视觉和人工智能的重点研究内容. 近年来, 随着视频获取方式的快速革新和拍摄需求的多样化发展, 视频数据的智能化分析任务面临着新的机遇与挑战, 涌现出众多的视频萃取方法. 本文创新性地从信息论的角度, 解释了数据、信息和知识之间的关系, 确立了视频萃取的核心是提高单位数据量的信息提供能力这一基本原则, 并依据数据信容 (information capacity) 分析, 从理论上对视频萃取中的各项任务进行了统一. 进一步地, 分类讨论了视频时空表征中的关键问题与解决方案, 系统地分析了从内容、目标和语义角度进行视频萃取的方法, 结合视频摘要、浓缩和描述任务, 梳理出三条发展主线, 展现了视频萃取的发展态势. 更重要的是, 本文对现有方法的优势与缺陷进行了深入的思考与讨论, 指出了尚未解决的若干关键科学问题, 并对解决方案进行了初步探讨. 同时, 本文对视频萃取研究所面临的挑战与未来发展趋势进行了系统的分析与展望.

关键词 视频萃取, 视觉表征, 视频摘要, 视频浓缩, 视频描述, 计算机视觉, 人工智能

1 引言

随着视频采集设备的普及, 视频数据呈现爆炸式增长, 成为人们日常生活中最重要的信息载体之一. 由于视频所承载的信息量丰富, 能够给人更为直观的视听感受, 视频数据在社会生活的各领域得到广泛应用, 成为最重要的信息载体之一 [1~3]. 同时, “互联网 +” 时代的到来, 进一步促使海量视频数据流向互联网. 由于缺乏对上传数据的有效监管, 各大网络视频平台 (如 YouTube, bilibili, 抖音等) 存在大量冗余和低质量的视频, 对用户快速获取有用的视频信息造成了极大的困扰 [4,5]. 因此, 如何有效地萃取视频中的主要信息, 去除信息增益较少的冗余、低质量和不相关的内容, 进而获得更为简洁的数据形式和信息模态, 是视频智能化分析的基础问题 [6~9]. 在此背景下, 视频萃取技术应运而生. 视频萃取 (video distillation) 通过研究视频数据的时空和语义特性, 探索简洁高效的数据展示形式和信息

引用格式: 李学龙, 赵斌. 视频萃取. 中国科学: 信息科学, 2021, 51: 695–734, doi: 10.1360/SSI-2020-0165
Li X L, Zhao B. Video distillation (in Chinese). Sci Sin Inform, 2021, 51: 695–734, doi: 10.1360/SSI-2020-0165

感知模态, 致力于提高单位数据量的信息提供能力, 是人工智能的关键技术之一. 视频萃取的形式多种多样, 如提取关键帧和关键镜头来概括视频内容, 分析前景目标运动轨迹来浓缩视频中的运动信息, 或者生成跨模态文本语句来描述视频中的活动等, 对海量视频数据的浏览、检索、存储等具有重要意义 [10~12]. 另外, 视频萃取对基于视频数据的其他人工智能技术, 如人机交互、辅助视觉、场景理解等具有重大的推动作用 [13]. 鉴于此, 世界知名的研究机构与公司, 如卡内基 · 梅隆大学 (Carnegie Mellon University)、麻省理工学院 (Massachusetts Institute of Technology)、FX Palo Alto 实验室、阿里巴巴达摩院与微软亚洲研究院等, 纷纷加入到视频萃取的研究中, 极大地推动了视频数据智能分析的相关技术发展.

实际上, 从数据变化的角度, 可以将基于视频数据的智能分析任务分为 4 类. 假设视频中原有的数据形式为 A.

• A $\rightarrow$ a: 这类任务将视频数据转变为语义标签的形式, 包含视频分类、目标检测、场景识别、行为分析等任务.

• A $\rightarrow$ A+: 这类任务通过对视频数据的先验信息进行建模, 将先验信息注入到原视频数据中, 提高原视频的信息量, 包含视频超分辨率重建、视频超帧率重建等任务.

• A $\rightarrow$ A′: 这类任务通过对视频数据进行后期处理, 在不影响原视频数据量的前提下转变数据形态, 包含视频去噪、风格转换、画面增强等任务.

• A$\rightarrow$ A–: 这类任务在理解视频内容的基础上, 对原视频内容进行提取、浓缩和跨模态转换, 以去除冗余, 也就是本文所说的视频萃取任务.

视频萃取的任务是从原始数据中提取出关键信息, 同时去除携带冗余信息和无意义信息的数据, 进而提高视频单位数据量的信息提供能力. 假设某一视频数据所提供的信息量为 I, 数据量为 D, 用信容 (information capacity, IC) 代表视频单位数据量的信息提供能力, 计算公式如下:

$$\psi = \frac{I}{D}, \tag{1}$$

信容 ψ 同时反映了视频的信息稠密度和数据紧致度. 显然, 视频萃取任务就是在保持视频信息量 I 的基础上, 尽可能地减少数据量 D, 进而提升视频信息提供能力的过程. 进一步地, 考虑到视频萃取的目的是为了让观看者高效地理解视频信息, 萃取内容的展示形式应该考虑人类的认知能力, 即观看者从视频中获取的知识, 取决于视频所提供的信息在观看者认知能力上的投影, 表示如下:

$$\kappa = \langle \boldsymbol{I}', \Lambda \rangle = \langle \Psi' \odot \boldsymbol{D}', \Lambda \rangle , \tag{2}$$

其中, κ 表示观看者从视频中获取的知识. $\boldsymbol{I}'$, $\boldsymbol{D}'$, Ψ' 和 Λ 分别代表视频萃取后信息量、数据量、信息提供能力, 以及人类认知能力的向量化表示. 这主要是考虑了视频内容和人类认知的多面性 [14, 15]. 本文中, 矢量加粗表示.

目前, 研究人员将视频萃取定义为一个多学科交叉的研究方向, 涉及计算机视觉、多媒体分析、机器学习、自然语言处理等多个研究领域. 结合视频数据的特点和人类理解的便捷性, 探索了视频摘要 (video summarization) [16]、视频浓缩 (video synopsis) [17] 和视频描述 (video captioning) [18] 3 个任务, 分别从内容、目标和语义角度进行视频萃取. 三者的关系如图 1 [19~62] 所示, 具体如下所述.

(1) 视频摘要从内容角度进行萃取. 它通过自动分析视频内容移除冗余信息, 并提取出最能代表原视频内容的关键帧或关键镜头, 从而达到从内容角度萃取视频信息的目的. 借助视频摘要, 人们可以快速定位感兴趣的内容, 提高视频浏览效率. 此外, 视频摘要还可作为其他视频分析任务的预处理过

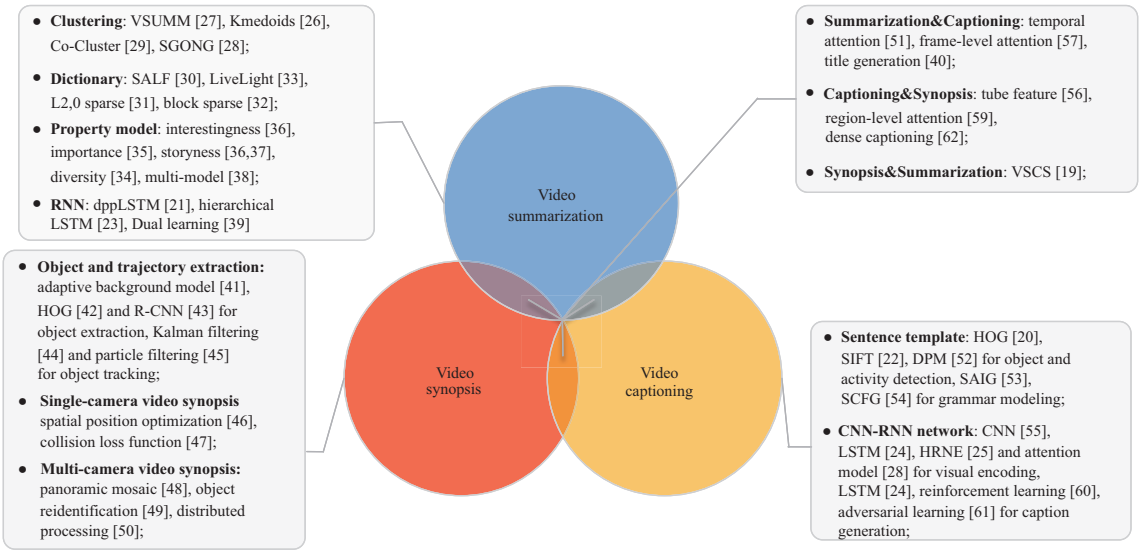

图 1 (网络版彩图) 视频萃取中, 摘要、浓缩和描述任务的关系图

Figure 1 (Color online) The relationship among video summarization, video synopsis and video captioning

程, 通过摘要得到更为简洁的视频数据, 可以显著提高场景理解、目标跟踪、行为识别等视频分析任务的算法效率 [63~65].

(2) 视频浓缩从目标角度进行萃取. 该任务主要针对监控视频开展研究, 通过对视频中目标的运动轨迹进行自动化提取与分析, 以空间换取时间的方式, 在保持视频前景目标原有表观和运动信息的前提下, 为原视频提供一种更加简短和紧凑的视觉呈现方式, 压缩视频数据量的同时降低了存储压力, 进一步方便了监控视频中活动目标的快速检索, 是实现智能安防、辅助刑侦和异常事件预警的重要手段 [17,19].

(3) 视频描述从语义角度进行萃取. 它同时利用计算机视觉和自然语言处理技术, 通过对视频内容进行语义层次的理解, 实现视觉信息到文本信息的跨模态转换, 最终生成文本语句来描述视频内容. 视频描述对行为分析、人机交互、计算机辅助视觉等应用具有重要意义 [66,67].

视频摘要、浓缩和描述任务是视频萃取研究的 3 条主线, 它们依托视频数据时空表征的研究基础, 分别以画面、像素和文本为任务出口, 从内容、目标和语义角度进行视频萃取. 近年来, 在科研人员的共同努力下, 视频萃取的相关研究已经取得了长足发展. 为了还原视频萃取研究的全貌, 本文在对视频数据表征方法进行介绍的基础上, 从视频摘要、视频浓缩和视频描述任务入手, 对相关方法进行了详细的讨论, 理清了历史发展沿革, 分析了现有方法的优势和缺陷, 指出了各任务在视频数据爆炸时代所面临的机遇与挑战, 并对该领域的未来研究趋势进行了展望.

本文的组织框架具体如下: 第 2 节, 介绍了视频萃取任务中常用的视频数据表征方法, 分为空间信息表征和时序信息表征; 第 3 节, 从聚类算法、字典学习、性质模型和递归神经网络 4 个方面, 介绍了视频摘要方法的历史发展沿革和研究现状, 并对现存问题进行了分析与讨论; 第 4 节, 介绍了视频浓缩任务中, 目标轨迹提取方法和基于单/多摄像头的视频浓缩方法, 并对现存问题进行了分析与讨论; 第 5 节, 分类介绍了基于语句模板和 CNN-RNN 框架的视频描述方法, 并对现存问题进行了分析与讨论; 第 6 节, 探讨了视频萃取所面临的挑战, 展望了未来的研究发展趋势; 第 7 节, 进行全文总结.

表 1 视频空间信息表征方法

Table 1 Video spatial feature extraction approaches

Feature type	Approach	Pretraining dataset
Hand-crafted feature	Color histogram [70], HOG [20], SIFT [22], SURF [71]	–
Deep learning feature	AlexNet [72], VGGnet [73], FCN [74], ResNet [75], GoogLeNet [76], DenseNet [77]	Image datasets: ImageNet [78], OpenImage [79], MNIST [80], CIFAR10 [81], Places [82], Video datasets: HMDB51 [84], UCF101 [83], Moment-in-Time [85], ActivityNet [86], Kenetics [87]

2 视频数据表征方法

视频数据表征是建模与分析的前提, 也是视频萃取的基础. 视频数据由图像序列组成, 兼具丰富的空间信息和时序信息, 下文将对视频数据的空间信息和时序信息表征方法进行分类介绍.

2.1 视频空间信息表征

视频空间信息表征主要沿用图像数据的表征方法, 可大致分为传统手工设计特征 (hand-crafted features) 和深度学习特征 (deep learning features). 近年来, 得益于深度神经网络模型强大的特征学习能力, 深度学习特征在各种视频分析任务中有全面超越手工设计特征的趋势. 但是, 手工特征基于人类对不同事物和任务的认知而设计, 具有较强的可解释性. 同时, 手工特征的设计经验可以启发深度神经网络的构造与学习, 因此它在视频表征中仍然发挥着至关重要的作用 [68, 69]. 表 1 [20, 22, 70~87] 总结了视频萃取中常用的手工设计特征和深度学习特征, 以及训练深度网络所需要的图像视频数据集. 下文将分别进行介绍.

2.1.1 手工设计特征

视频数据的手工特征可以分为全局特征 [70] 和局部特征 [88, 89] 两种. 视频萃取任务中, 常见的视频帧全局特征有颜色直方图特征 (color histogram) [70] 和方向梯度直方图特征 (HOG) [20] 等. Song 等 [90] 和 Zhang 等 [21] 将颜色直方图特征利用到视频摘要任务中. 颜色直方图是对图像中颜色组成的特征统计, 它将颜色空间的每个通道均匀分割为若干个桶 (bin), 然后将图像中的每个像素点按照区间值分配到不同的桶中, 进而统计每个桶中像素点出现的频率, 并将其作为颜色特征的一维, 最后所有桶的统计值连接, 获得颜色直方图的特征向量. 颜色直方图特征通常针对 RGB 和 HSV 颜色空间进行提取, 以 RGB 空间为例, 若将 R, G, B 三个通道切分的桶个数分别记为 m, n, p, 那么最终颜色直方图的维度为 $m \times n \times p$. 颜色直方图的优点是只统计不同颜色出现的频率信息, 简单高效, 且具有一定的旋转不变性, 缺点是对光照变化敏感, 且忽略了颜色的空间位置信息 [91].

Dalal 等 [20] 于 2005 年提出方向梯度直方图 HOG (histogram of oriented gradient). HOG 是用于统计图像梯度方向和梯度强度分布的特征描述子, 在目标跟踪 [92]、行为识别 [93]、显著性检测 [94] 等视频分析任务中具有广泛的应用. 在提取 HOG 特征时, 首先需要将图像灰度化, 并进行亮度值的 Gamma 矫正, 然后对图像中的每个像素点计算梯度值与方向, 最后分块统计像素点的梯度方向, 并利

用梯度值对每个方向进行归一化, 获得特征向量. 方向梯度直方图对图像几何和光学形变具有良好的鲁棒性, 但是计算速度较慢, 并且对图像噪点较为敏感, 不适用于低画面质量的视频数据.

局部特征是一种表征图像小块区域或者特征点的特征类型. Lowe [22] 于 1999 年提出 SIFT (scale-invariant feature transform). 在众多计算局部特征的方法中, SIFT 是一种常用的视频空间特征. SIFT 通过计算图像中局部特征点的尺度与方向描述子从而得到特征向量, 目前广泛应用于视频摘要 [23]、检索 [95]、稳像 [96] 和事件检测 [97] 等任务中. 首先, SIFT 通过对原图像进行尺度变换, 构建图像尺度空间, 进而得到高斯差分金字塔, 这一步的目的是为了对不同分辨率下的图像进行特征提取. 其次, 在此基础上, 针对图像金字塔进行空间极值点检测, 获取稳定极值点的位置信息, 进而计算极值点邻域内的梯度幅值和方向. 最后, 统计图像极值点的梯度分布, 生成梯度方向直方图作为图像特征向量. SIFT 的优点是对图像的旋转平移、视角切换、大小缩放、亮度变化等具有较强的稳定性, 缺点是计算复杂度高、速度较慢. 为了进一步提高 SIFT 特征的计算效率, 2006 年 Bay 等 [71] 提出了 SURF (speeded-up robust features), 采用箱式滤波器对图像进行处理, 利用 Hessian 矩阵构造金字塔尺度空间, 避免了 SIFT 特征中对图像进行降采样的操作, 提高了算法效率. 总体来说, SURF 特征对于图像尺度和旋转变换的鲁棒性不及 SIFT 特征, 但是其速度是 SIFT 的三倍.

2.1.2 深度学习特征

近年来, 随着深度学习的发展, 利用深度卷积神经网络 CNN (convolutional neural networks) 进行视频空间特征提取的方法陆续出现. 不同于认知驱动的手工设计特征, 深度特征是在数据驱动下进行自主学习的特征类型. 视频萃取任务中常用于提取深度特征的 CNN 网络有 AlexNet [72], VGGnet [73], FCN [74], ResNet [75], GoogLeNet [76], DenseNet [77] 等, 具体如下.

Krizhevsky 等 [72] 于 2012 年提出 AlexNet. AlexNet 具有 5 个卷积层和 3 个全连接层, 总训练参数量达到 6.24×10^8, 其中全连接层的参数占据了 94.2%. 得益于卷积神经网络强大的视觉特征提取能力, AlexNet 在当年 ImageNet 的图像分类比赛中极大地超越了当时最优的传统方法. Simonyan 等 [73] 于 2013 年提出 VGGnet. VGGnet 包括 VGGnet-16 和 VGGnet-19 两种网络结构, 其中 VGGnet-16 包含 13 个卷积层和 3 个全连接层, VGGnet-19 包含 16 个卷积层和 3 个全连接层. 相比于 AlexNet, VGGnet 最大的改进是利用小卷积核代替大卷积核, 比如利用 2 层 3×3 卷积核来代替 5×5 卷积核, 利用 3 层 3×3 卷积核来代替 7×7 卷积核等, 在保持网络层感受野尺度的前提下, 增加了网络深度, 相应地提高了网络的学习能力. 但是, VGGnet 和 AlexNet 都存在全连接层参数量过多的问题. 针对此问题, Long 等 [74] 在 2015 年提出了全卷积神经网络 FCN (full convolutional neural networks), 利用 1×1 卷积全面替代全连接操作, 显著减少了模型参数, 同时克服了传统 CNN 中全连接层对图像尺寸的限制. 另外, 为了进一步增强 CNN 的学习能力, He 等 [75] 和 Szegedy 等 [76] 分别提出了 ResNet 和 GoogLeNet, 用于拓展 CNN 网络的深度和宽度. ResNet 在卷积层之间添加了跳连操作来构建残差模块 (residual blocks), 克服了网络层数加深导致的模型学习梯度消失和爆炸的问题, 拓展了 CNN 的深度. 与 ResNet 不同, GoogLeNet 致力于拓展 CNN 的宽度, 为此设计了 Inception 模块, 在每一层利用 4 个不同尺寸的卷积核平行处理图像, 并进行聚合. Inception 模块拓展了 CNN 的宽度, 能够更高效地利用 GPU 计算资源, 在相同的计算量下能提取更多的特征, 从而提高网络学习能力. 实际上, GoogLeNet 的参数量只是 AlexNet 的 1/12. 此外, 为了兼顾 CNN 的浅层信息和深层信息, DenseNet 在任意两个卷积层之间加入了密集连接机制, 改变了传统 CNN 只在相邻的卷积层之间进行连接的结构, 进一步提高了 CNN 的学习能力.

上述 CNN 网络在提取视频数据的深度特征时, 首先需要在大型图像或视频分类数据集上进行预

表 2 深度学习预训练数据集

Table 2 Pretraining datasets for deep learning

Dataset type	Name	Size	Category	Year
Image dataset	MNIST [80]	60000	10	1998
	ImageNet [78]	120 million	1000	2009
	CIFAR10 [81]	60000	10	2009
	Places [82]	10 million	434	2014
	OpenImage [79]	900 million	6000	2016
Video dataset	HMDB51 [84]	6766	51	2011
	UCF101 [83]	13320	101	2012
	ActivityNet [86]	20000	200	2015
	Moment-in-Time [85]	903964	339	2018
	Kenetics [87]	500000	600	2018

表 3 视频时序信息表征方法

Table 3 Video temporal feature extraction approaches

Feature type	Approach
Flow feature	Optical Flow [104], FlowNet [105], FlowNet2.0 [106], SPyNet [107], PWCNet [108], EpicFlow [109]
3D convolution feature	3D CNN [110], C3D [111], I-3D [112], ResNet-3D [113], Pseudo-3D [114]
RNN feature	LSTM [115], GRU [116], Hierarchical LSTM [23, 25], HSA-LSTM [117], MA-LSTM [118]

训练, 用于学习数据与标签之间复杂的非线性映射 [98, 99]. 如表 2 所示, 常用的图像预训练数据集有 ImageNet [78], OpenImage [79], MNIST [80], CIFAR10 [81], Places [82] 等, 常用的视频预训练数据集有 HMDB51 [84], UCF101 [83], Moment-in-Time [85], ActivityNet [86], Kenetics [87] 等. 预训练的 CNN 将直接应用于不同的视频分析任务中进行特征提取 [21, 100], 或者在相应的任务中进行参数微调以及模型迁移 [101, 102]. 借助于 CNN 强大的学习能力, CNN 在大型数据集上学习得到的深度特征在挖掘图像和视频的高层语义信息方面具有更为优越的性能 [103], 因此深度特征被应用于各种各样的视频分析任务中, 基于深度特征的视频萃取方法也往往能取得相对于传统手工特征更好的结果.

2.2 视频时序信息表征

视频是由连续变化的图像组成的, 相对于空间二维图像, 视频具有独特的时序特性. 视频时序表征的重点在于挖掘视频的运动、变化和序列信息, 主要包含光流运动特征、三维卷积神经网络特征和递归神经网络特征. 表 3 [23, 25, 104~118] 统计了视频萃取中常用的时序特征, 下文将具体分类介绍.

2.2.1 光流运动特征

Gibson [104] 在 1950 年首先提出光流的概念, 用于对视频中的运动信息进行描述. 光流特征 (optical flow) 是视频萃取中常用的时序特征 [23, 24, 119], 其通过计算相邻视频帧的空间相关性以及每个像素在

时间上的变化, 确定相邻帧之间的对应关系, 记录了视频中相邻帧间像素级的相对运动信息. 传统的光流特征计算方法主要包含基于梯度、多项式和区域匹配的方法等. 值得注意的是, 传统的光流计算方法需要满足两个基本假设: (1) 在相邻帧中, 目标的亮度没有发生改变; (2) 目标运动连续且位移较小. 这限制了传统光流特征对不同场景的适用性, 而且计算复杂度较高, 影响视频萃取效率.

为了进一步提高光流特征的适用性和计算效率, Dosovitskiy 等 [105] 在 2015 年提出了基于卷积神经网络的光流计算方法, 即 FlowNet. FlowNet 采用 Encoder-Decoder 的网络结构, 以利用人工合成数据集的光流图作为监督信息, 以相邻的两个视频帧作为输入图像, 输出得到光流图, 实现了视频帧到光流图的端到端学习. FlowNet 可以进行光流图的实时计算, 并在实时的光流估计算法中取得了最优的结果, 但是 FlowNet 的光流计算精度要明显低于传统方法. 基于此, FlowNet 团队在 2017 年提出了改进的 FlowNet2.0 版本 [106], 采用由粗糙到精细 (coarse-to-fine) 的光流估计策略, 通过堆叠多个 FlowNet 网络, 在保持实时性的基础上实现了光流特征的高精度提取. 近几年, 在 FlowNet 和 FlowNet2.0 的启发下, 越来越多的光流卷积神经网络被提出, 如 SPyNet [107], PWCNet [108], EpicFlow [109] 等, 广泛应用于视频超帧率重建、行为识别和场景分类等智能分析任务中.

2.2.2 三维卷积神经网络特征

Ji 等 [110] 于 2010 年提出三维卷积神经网络 (3D CNN), 广泛应用于行为识别、视频分类等任务的特征提取. 3D CNN 利用三维卷积核替代二维卷积核, 同时在空间和时间维度对连续多个视频帧进行数据处理, 可以捕捉视频流的时序信息. 实际上, 3D CNN 在对视频片段进行处理的过程中, 随着时间尺度的增加, 模型参数也相应地增加. 由于计算量较大, 同时受限于硬件设备的计算能力, 3D CNN 只能用于短视频片段 (7 帧以内) 的时序特征提取 [110].

在 3D CNN 的启发下, 研究人员结合成熟的二维卷积神经网络框架, 提出了 C3D [111], I-3D [112], ResNet-3D [113], Pseudo-3D [114] 等三维卷积神经网络用于视频时序特征的提取. 这些方法都是视频分析任务中常用的时序特征提取方法, 但是它们都存在一个共同的问题: 无法对视频帧序列的长时依赖关系进行建模, 例如 C3D 只能处理连续 16 帧的序列. 然而, 视频往往具有成千上万帧, 这在一定程度上限制了三维卷积神经网络所提取的时序特征对视频分析任务的适用性.

2.2.3 递归神经网络特征

鉴于递归神经网络 RNN (recurrent neural networks) 强大的时序建模能力, 及其在机器翻译、文本分类等时序分析任务中取得的巨大成功, 研究人员也将 RNN 引入到视频时序特征提取任务中 [21,120]. 传统 RNN 是通过对前向网络增加反馈连接建立而成的, 所以它具有递归处理序列数据的能力. 但是, 传统 RNN 在处理序列数据时容易出现梯度爆炸和消失的问题, 导致网络非常难以训练 [121]. 为解决此问题, Hochreiter 等 [115] 提出了长短时记忆网络 LSTM (long short-term memory), 通过为 RNN 增加 3 个门限结构实现模型学习时的梯度稳定, 现已成为 RNN 最流行的变体. GRU(gated recurrent unit) [116] 也是 RNN 一种常见的变体, 和 LSTM 具有相近的性能, 但是结构更为简单, 计算效率更高, 常用于视频分类和描述等任务中.

相比于三维卷积神经网络, 递归神经网络可以对相对较长的视频片段进行时序特征提取. 实验表明, LSTM 可以对 100 帧以内的短视频片段进行有效的时序特征提取 [23~25]. 但是, 视频萃取任务中的视频序列往往长度较大. 一般来说, 用于摘要的视频至少具有数分钟的时长, 按照 25 帧/秒计算, 每个视频包含了成千上万个视频帧. 现有方法很难对这种长度量级的视频序列进行有效的时序特征提取. 基于此, Pan 等 [25] 和 Zhao 等 [23] 分别针对视频描述和视频摘要任务提出了层次化的递归神经网络

模型 (hierarchical LSTM), 只需要较少的计算单元就可以实现长时视频的时序信息建模, 拓展了单层 LSTM 的时序建模能力, 还可以进一步提高 LSTM 的非线性拟合能力. 在此基础上, Zhao 等 [117] 进一步提出了结构自适应的递归神经网络 (hierarchical structure-adaptive LSTM, HSA-LSTM), 在时序特征提取时考虑视频结构信息的影响, 显著提高了视频时序特征的判别能力. 此外, Arabshahi 等 [118] 挖掘了递归神经网络的记忆力增强机制 (memory-augmented LSTM, MA-LSTM), 通过加入额外的记忆单元增强递归神经网络对时序信息的记忆能力, 进一步提高了传统递归神经网络对长视频时序特征的提取能力.

3 视频摘要

如图 2 所示, 视频摘要旨在从冗长的视频序列中提取出关键帧或关键镜头, 进而概括原视频的主要内容. 利用视频摘要技术可以显著提升信息检索的效率, 方便观众进行视频浏览, 另外也从数据压缩的角度缓解了海量视频数据的存储压力.

本文尝试从信息论的角度对视频摘要任务进行建模. 给定视频帧序列 $\boldsymbol{V}=\{f_1, f_2, \ldots . f_n\}$, 视频摘要的目的是从中挑选出子序列 $\boldsymbol{S}=\boldsymbol{V}\cdot\boldsymbol{L}$ 来代表原视频的主要内容, 其中 $\boldsymbol{L}\in\mathbb{R}^n$ 每一维的取值为 $\{0,1\}$, 代表对应位置视频帧是否被选入摘要中. 进一步地, 原视频中单位数据量的信息提供能力表示如下:

$$\varphi\left(\boldsymbol{V}\right)=\frac{I\left(\boldsymbol{V}\right)}{D\left(\boldsymbol{V}\right)}=\frac{I\left(f_1, f_2, \ldots . f_n\right)}{\sum_{i=1}^{n} D\left(f_i\right)}. \tag{3}$$

需要注意的是, 考虑到视频画面变化的连续性, 许多视频帧是非常相似的 (尤其对于相邻帧), 而相似帧往往无法带来额外的信息增益. 因此, 视频的信息量并不是每一帧信息量的简单加和, 即

$$I\left(f_1, f_2, \ldots . f_n\right)\neq\sum_{i=1}^{n} I\left(f_i\right). \tag{4}$$

假设视频 $\boldsymbol{V}$ 只包含 f_1 和 f_2 两帧, 它们作为整体所提供的信息量应该为两者信息量之和减去两者的互信息, 即

$$I\left(f_1, f_2\right)=I\left(f_1\right)+I\left(f_2\right)-I\left(f_1; f_2\right). \tag{5}$$

根据上式, 视频数据的通用信息量计算方法可归纳为

$$I\left(\boldsymbol{V}^{t+1}\right)=I\left(\boldsymbol{V}^{t}\right)+I\left(f_{t+1}\right)-I\left(\boldsymbol{V}^{t}; f_{t+1}\right), \tag{6}$$

其中 $\boldsymbol{V}^{t+1}=\{f_1, f_2, \ldots, f_{t+1}\}$, $\boldsymbol{V}^{t}=\{f_1, f_2, \ldots, f_t\}$.

视频摘要过程就是提升原视频单位数据量信息提供能力的过程, 即

$$\max_{\boldsymbol{L}}\varphi\left(\boldsymbol{S}\right)=\frac{I\left(\boldsymbol{S}\right)}{D\left(\boldsymbol{S}\right)}=\frac{I\left(\boldsymbol{V}\cdot\boldsymbol{L}\right)}{\sum_{i=1}^{n} D\left(f_i\right)\cdot l_i}, \tag{7}$$

其中, $l_i\in\{0,1\}$ 代表 $\boldsymbol{L}$ 的第 i 个元素. 由式 (6) 可知, 每增加一个关键帧或者关键镜头, 给摘要带来的信息增益为

$$\nabla I\left(\boldsymbol{S}^{t+1}\right)=I\left(k_{t+1}\right)-I\left(\boldsymbol{S}^{t}; k_{t+1}\right), \tag{8}$$

其中, $\boldsymbol{S}^{t+1}=\{k_1, k_2, \ldots . k_{t+1}\}$, $\boldsymbol{S}^{t}=\{k_1, k_2, \ldots . k_t\}$, k 代表关键帧或关键镜头. 由上式看出, 在摘要过程中, 应尽量挑选互信息少的关键帧或关键镜头集合, 以此使式 (7) 取得最大值.

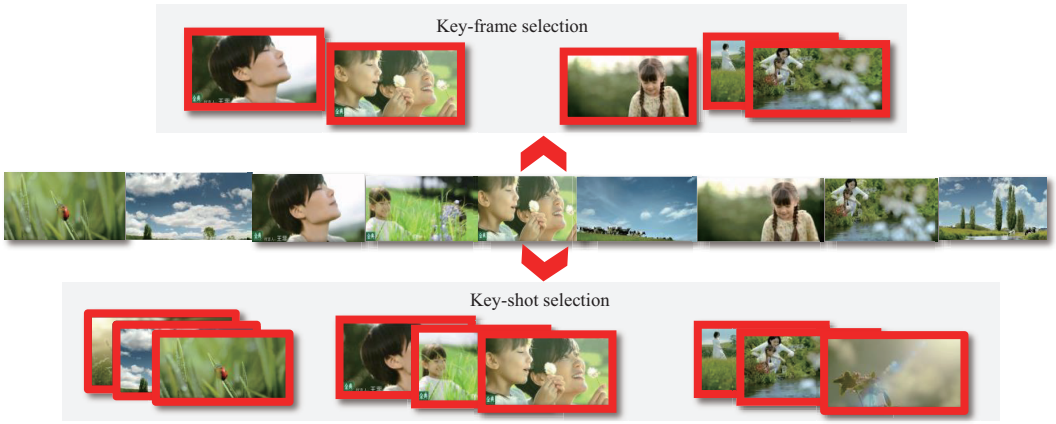

图 2 (网络版彩图) 视频摘要示意图

Figure 2 (Color online) Demonstration of video summarization

为了从内容角度进行视频萃取, 研究人员提出了多种多样的视频摘要方法. 按照所提出的时间顺序, 视频摘要任务大致分为 4 个发展阶段, 代表方法分别为基于聚类算法 (clustering)、字典学习 (dictionary learning)、性质模型 (property model) 和递归神经网络 (RNN) 的方法. 接下来, 本节将分别进行介绍.

3.1 基于聚类算法的视频摘要

聚类算法假设样本点集合分布在若干个聚类中心的周围, 通过特定的度量方式确定样本点之间的相似性, 并将相似的样本点分配到对应的类簇中 [122,123]. 基于聚类算法的视频摘要方法利用 K-means [124], K-medoids [26], Affinity Propagation [125], Density Peak [126] 等, 将视频帧或镜头分配到不同的类簇中, 通常最靠近聚类中心的帧或镜头被作为关键帧和关键镜头选入到视频摘要中.

上述方法将视频帧看作是离散的样本点, 直接将传统的聚类算法应用于视频摘要任务, 是聚类算法在视频摘要任务中的简单尝试, 并没有考虑视频摘要的数据特点, 因此算法性能较低. 视频数据的一个典型特点是视频帧具有时序变化的连续性. 基于此, Avila 等 [27] 提出了基于 K-means 的 VSUMM 方法, 对视频帧的类簇进行时序初始化, 相对于随机初始化的方法取得了更好的结果, 同时提高了 K-means 算法的收敛速度. 类似地, Ren 等 [127] 提出了基于视频中活动信息的镜头分割方法, 并按照时间顺序对镜头进行聚类. Papadopoulos 等 [28] 利用 SGONG 网络对视频镜头聚类, 该方法的优点是可以利用视频结构信息自动确定聚类的类簇数目, 减少摘要过程的人工干预, 这是对视频摘要方法的一次重大改进.

大多数基于聚类算法的视频摘要方法都只关注于聚类中心的计算, 却忽略了聚类中心的视觉质量. 实际上, 关键帧的画面、色彩和物体都严重影响观众对视频内容的理解. 为了进一步提升关键帧的质量, Khosla 等 [128] 在 K-means 聚类算法的基础上, 引入了网络图像的视角先验信息, 以保证所挑选的关键帧具有更多的信息量. 此外, 为了进一步提高视频摘要的效率, Chu 等 [29] 认为具有相同主题的视频往往具有类似的内容, 基于此提出了视频主题相关的联合聚类算法, 依靠不同视频间的视觉相似度进行摘要的联合生成. 该方法巧妙地利用了统计学习的思想, 可以一次性为具有相同主题的多个视频生成摘要. 但是, 该方法需要视频的主题标签, 然而大部分视频的主题并不明确, 这在一定程度上限制了其适用性.

3.2 基于字典学习的视频摘要

该类方法将视频摘要看作是视频内容的一个简单字典, 摘要的质量由字典对视频内容的重建能力决定, 重建能力由重建误差来衡量, 所以关键镜头的提取问题就转变为字典元素的选择问题 [129, 130]. 基于字典学习的视频摘要方法可以分为两种, 分别为基于稀疏自重建的方法和在线字典学习的方法.

基于稀疏自重建的方法利用视频镜头的子集对整体视频进行重建, 重建贡献高的镜头则表达能力强, 被选入视频摘要中 [30, 32, 129]. 区别于传统的字典学习方法, 研究人员在视频摘要任务中的创新主要集中在先验信息的建模上, 例如, Mei 等 [31] 根据视频摘要的任务特点, 提出了基于 $l_{2,0}$ 范数的字典学习方法, 更好地保证了关键帧的稀疏性. Wang 等 [131] 在稀疏字典学习的基础上加入了相似性抑制约束, 确保所挑选关键帧之间的差异性, 减少了视频摘要中的信息冗余. 这是传统字典学习方法忽略的问题, 但是对视频摘要的质量至关重要. 进一步地, 为了利用相邻帧具有相似性的先验信息, Ma 等 [32] 设计了基于块稀疏字典学习的视频摘要方法, 保持字典权重的局部相似性, 这样可以加快算法的收敛速度, 也相应地提高了性能. 结合聚类和字典学习算法, Zhao 等 [14] 提出了摘要空间的概念, 首先利用聚类算法获得字典元素, 然后在流行结构保持的约束下构建摘要空间, 度量每个视频帧对各字典元素的表达能力, 进而进行关键帧的挑选. 为了实现对视频摘要过程的全局优化, Etezadifar 等 [132] 提出了字典学习与关键帧挑选的联合优化框架, 提高了视频摘要的质量. 总体而言, 字典学习算法的目的是从视频帧集合中找到最能代表原视频内容的关键帧.

为了实现视频摘要任务的在线进行, 研究人员提出了在线字典学习的方法 [33, 65]. 这类方法首先将视频开头的部分作为初始字典, 按照时间顺序对视频内容进行重建, 重建误差较大的镜头说明其内容相对于当前字典较为新颖, 所以被选为新的字典元素, 依此进行视频摘要的在线生成. 类似地, Zhang 等 [65] 提出了一种在线的自动编码器, 模拟在线字典学习的方式, 可以同时进行特征学习和字典更新, 自动确定关键镜头的位置. 进一步地, Marvaniya 等 [133] 联合使用字典学习、全局相机运动分析和视频帧的色彩信息, 对每个视频帧被选入摘要的概率进行预测, 提高了视频摘要的质量, 并实现了视频摘要在移动设备上的实时生成. 此外, Wang 等 [134] 专门针对无人机拍摄的视频进行摘要, 依据此类视频场景不断变化的特点, 提出了一种在线场景分割的方法, 并在此基础上利用场景内视频帧的视觉显著性指标进行关键帧的挑选. 总而言之, 在线字典学习的方法为视频摘要的实时生成提供了新的思路, 使视频拍摄和摘要的同步进行成为了可能.

3.3 基于性质建模的视频摘要

为了准确地萃取视频信息, 视频摘要需要满足若干基本性质 [135], 比如: 包含原视频中的重要物体、代表原视频的主要内容、尽量少的信息冗余, 以及保持视频情节的流畅性等. 上述性质对视频摘要的质量至关重要. 如何对视频摘要的各项性质进行度量, 进而探索各性质之间的关系, 是基于性质建模的视频摘要方法的研究重点 [36, 38].

现有方法提出不同的性质模型对生成摘要进行约束, 主要包含差异性、代表性、重要性、情节性等模型. 首先, 为了有效地去除冗余信息, 相同或相似的镜头不可以同时出现在视频摘要中 [136, 137]. 现有方法设计了不同的度量方式来计算视频特征向量之间的相似性. 例如, Ren 等 [127] 利用颜色直方图之间的 Sogeral 距离来度量视频帧之间的相似性, 而 Ejaz 等 [138] 通过多种度量方式的线性组合进行差异性的度量. Gong 等 [34] 在改进行列式点过程模型 (determinantal point process, DPP) [139] 的基础上, 提出一个基于时序建模的 DPP 模型来保证关键帧的多样性. 为了有效地概括视频内容, 原视频内容关键帧和关键镜头还需要具有代表性. 前文所述的基于聚类和字典学习的方法是度量视频摘要代表

性常用的典型方法. 进一步地, 为了保证摘要中包含原视频中的重要物体, 重要性模型也逐渐被提出. 针对视频摘要重要性的建模主要是基于目标的方法, 将视频摘要问题转变为重要物体的检测问题 [35]. 基于目标的方法主要分为两类, 其中一类方法对重要物体进行了预先的定义, 并以此作为关键镜头的必要约束 [140]. 重要物体可以是单独的个体, 也可以是包含众多个体的群组 [141]. 另外一类方法通过前背景分离模型获取目标像素, 然后利用目标到视频帧中心的距离、出现的频率等区域特征训练回归模型, 进而预测每一目标区域的重要程度 [142, 143]. 最近, 深度神经网络也被用于视频镜头的重要性预测 [135]. 视频摘要的情节性模型要求摘要准确反映出原视频的故事情节发展, 而视频中的物体是推动情节发展的重要元素 [36]. 因此, 相邻两个关键镜头中出现的物体需要存在紧密联系, 这种联系通常通过概率模型度量 [37].

为了综合度量视频摘要的性质, 受摘要性质建模思想的启发, Gygli 等 [36] 同时设计了基于重要目标的趣味性模型、基于聚类算法的代表性模型和基于关键帧时序分布的情节性模型, 进而将各模型转换为子模函数形式, 利用人工摘要的监督信息进行协同优化, 在针对自拍视频的摘要任务中取得了巨大的成功. 类似地, Li 等 [38] 建立了剪辑视频 (edited video) 和未剪辑视频 (raw video) 的联合学习框架, 通过分析不同类别视频及其摘要的数据特点, 多方面利用视频信息, 设计了具有通用性的视频摘要重要性、代表性、差异性和情节性模型, 进而实现了不同性质模型的融合, 显著提高了视频摘要的质量. 总体而言, 相比基于单一性质模型的方法, 基于多性质模型联合优化的方法往往能取得更优的视频摘要性能.

3.4 基于递归神经网络的视频摘要

近年来, 得益于 RNN 强大的时序建模能力, 其在处理序列问题中取得了重大的进步 [144~146]. 由于视频数据是由图像序列组成的, 科研人员也将 RNN 引入到视频摘要任务中 [147, 148]. 这类方法以人工摘要作为监督信息, 通过 RNN 挖掘视频帧或镜头间的时序依赖关系, 进而预测每个视频帧和镜头的分值, 使得人工摘要和生成摘要具有最大的重合度. 其中, Zhang 等 [148] 利用双向 LSTM 来预测每一个视频镜头被选进摘要的概率. 另外, 为了保证关键镜头的差异性, 它引入了 DPP 模型来抑制相似镜头被同时选中的可能性. 考虑到单层 LSTM 只能对较短的视频序列进行建模, Zhao 等 [23] 在多层卷积的启发下设计了 LSTM 的分层结构. 分层 LSTM 共包含两层, 第 1 层由一个单向 LSTM 构成, 用于挖掘镜头内的帧间依赖, 并进行镜头特征的编码. 第 2 层是一个双向的 LSTM, 用于挖掘镜头间的前后向时序依赖关系, 并将其输出到每一步的隐状态, 最后进行关键镜头的挑选. 实验证明, 分层 LSTM 拓展了单层 LSTM 的时序建模能力, 还可以提高 LSTM 的非线性拟合能力. 进一步地, Zhao 等 [117] 提出了一种结构自适应的视频摘要方法, 该方法可以同时进行视频镜头分割和关键镜头挑选. 它弥补了现有方法在视频结构挖掘中的不足, 进一步提高了视频摘要的质量. 在这类算法中, RNN 主要用于挖掘视频数据中的时序依赖关系, 它们的主要贡献和创新点集中在基于视频数据特点的 RNN 结构设计上.

上述方法大都依赖于人工摘要的监督信息, 仅仅利用简单的判别损失函数进行模型优化. 这会导致模型训练时的指导信息不足, 增加模型学习难度. 为了给基于 RNN 的视频摘要模型提供更多的指导信息, 研究人员设计了不同的模型优化策略. 其中, Mahasseni 等 [149] 在 RNN 时序建模框架的基础上, 提出了对抗学习模型, 设计对抗网络辨别生成摘要的真假性. 受视频描述任务的启发, Apostolidis 等 [150] 在对抗学习的基础上, 引入了注意力模型来对视频帧和镜头特征进行自适应编码, 增强了特征的判别能力, 也相应地提高了算法性能. 类似地, Ji 等 [151] 建立了基于注意力模型的编码 – 解码网络实现关键镜头的自动化挑选.

表 4 视频摘要数据集统计

Table 4 Statistics on video summarization datasets

Dataset	Year	Size	Average duration	Video type	Summary type
YouTube [27]	2011	50	1∼10 min	Edited video	Key-frame
OVP [27]	2011	50	1∼4 min	Edited video	Key-frame
TVsum [90]	2015	50	2∼10 min	Edited Video	Key-shot
CoSum [29]	2015	51	about 3 min	Edited video	Key-shot
MED [154]	2014	160	1∼5 min	Edited video	Key-shot
SumMe [155]	2014	25	1∼6 min	Raw video	Key-shot
VTW [156]	2016	2529	about 1.5 min	Raw video	Key-shot
UTE [40]	2012	4	3∼5 h	Egocentric video	Key-shot

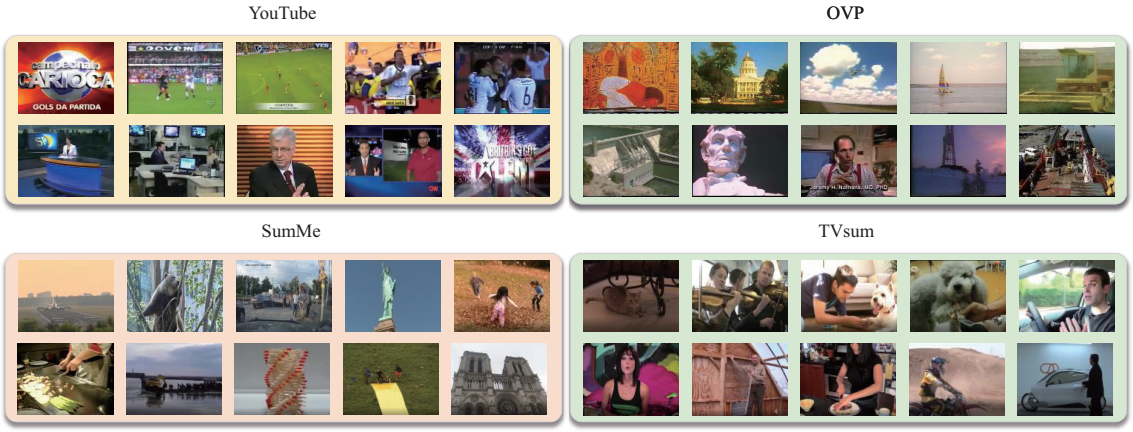

图 3 (网络版彩图) 视频摘要数据集展示

Figure 3 (Color online) The snapshots of video summarization datasets

此外, 由于人工摘要的挑选过程费时费力, 且具有一定的主观性, 所以现有的视频摘要数据集体量较小. 为了缓解模型学习对人工摘要的依赖, Rochan 等 [152] 将视频摘要看作序列标注问题, 设计了全卷积序列网络对视频的表观和时序信息进行联合建模, 并在此基础上开发了非配对优化策略, 在一定程度上缓解了视频摘要标注数据难以获取的问题 [153]. 为了给模型学习过程提供更多的指导信息, Zhou 等 [100] 在以 LSTM 作为摘要生成器的基础上, 借用传统方法中的代表性和差异性模型作为奖励函数, 构建摘要生成器的强化学习策略, 利用性质模型的反馈奖励辅助人工摘要的监督信息进行模型学习. 类似地, Zhao 等 [39] 建立了视频重构和摘要任务的对偶学习框架. 其中, 视频重构器可以为摘要生成器提供反馈信息, 指导摘要生成过程, 同时保证了生成摘要对原视频内容的重构能力. 这种自监督的方式为视频摘要模型的无监督学习提供了新的思路.

3.5 视频摘要数据集和评价指标

视频摘要任务常用的公开数据集如表 4 [27, 29, 40, 90, 154∼156] 所示, 部分视频在图 3 中进行展示. 按照视频类型划分, YouTube, OVP, TVsum, CoSum 和 MED 都是由经过人工剪辑的视频组成的数据集, 主要包含体育、烹饪、旅游、新闻、纪录片等视频类别. SumMe, VTW 和 UTE 是由未经过人工剪辑的视频组成的数据集, 它们主要来自于 GoPro、手机、智能眼镜等自拍设备. 按照摘要类型划分,

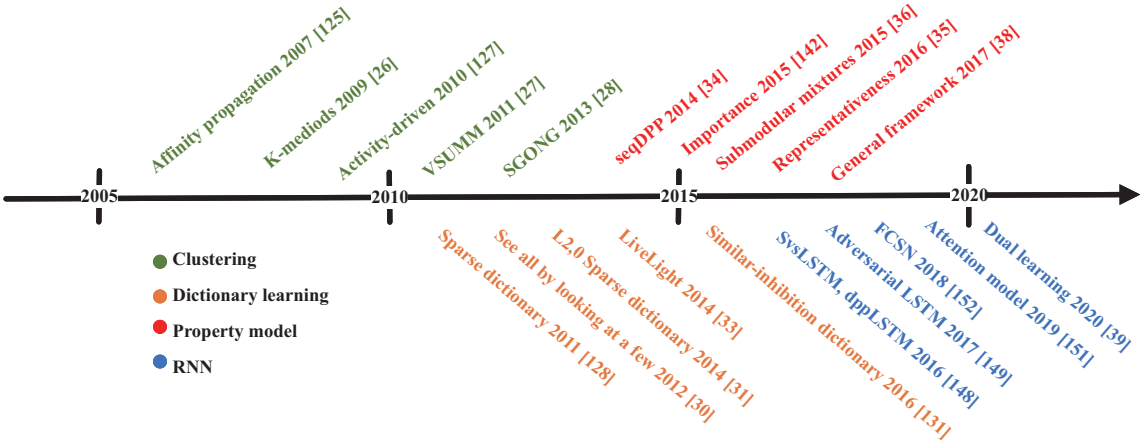

图 4 (网络版彩图) 视频摘要发展趋势图

Figure 4 (Color online) Development of video summarization

YouTube 和 OVP 是基于关键帧的视频摘要数据集, 其他都为基于关键镜头的视频摘要数据集, 这类视频数据集越来越多主要是因为基于关键镜头的视频摘要能够保持原视频的动态特性, 具有更好的可读性. 可以看出, 大部分用于视频摘要任务的视频时长在 1~10 min. 特别地, UTE 数据集中的视频长度约为 3~5 小时, 这是来自于谷歌智能眼镜的视频, 称为第一视角视频 (egocentric video), 它记录了佩戴者一段时间的连续活动, 具有大量的冗余信息和低质量片段, 针对这类视频的摘要任务对人们日常生活的记录与检索非常重要.

以上视频摘要数据集中大都具有人工标注的摘要标签. 不同算法生成视频摘要的质量由生成摘要与人工摘要的相似度来度量, 最常用的评价指标为准确率 (precision)、召回率 (recall) 和 F-measure. 它们的定义如下:

$$P=\frac{|S\cap S^{gt}|}{|S|},\quad R=\frac{|S\cap S^{gt}|}{|S^{gt}|},\quad F=2\cdot\frac{P\cdot R}{P+R},\tag{9}$$

其中, S 和 S^{gt} 分别代表算法生成摘要和人工标注摘要, F-measure 是准确率和召回率的调和平均值. 生成摘要 S 和人工摘要 S^{gt} 越相似, 上述 3 个指标分数越高, 也说明生成摘要的质量越高.

3.6 分析与讨论

视频摘要任务从内容角度进行视频萃取, 如图 4 所示, 经过数十年的发展, 视频摘要的主流方法已经从基于聚类和字典学习的方法, 逐渐过渡到基于高层语义的方法, 包括基于性质模型和深度学习的方法. 特别地, 得益于卷积神经网络强大的视觉特征提取能力和递归神经网络的时序建模能力, 基于深度学习的视频摘要方法取得了最优的性能, 相信短期内深度学习仍然在视频摘要任务中占据主导地位. 此外, 越来越多的研究工作也开始意识到传统方法的优势, 比如: 可解释性强、无监督学习等, 深度学习模型与传统算法相结合的方法也将是视频摘要任务发展的主流趋势.

近年来, 虽然视频摘要已经取得了重大发展, 但是仍然存在一些关键问题亟待解决, 具体如下.

(1) 视频摘要模型的无监督学习问题. 现有性能优越的视频摘要方法大都是基于有监督学习的方法, 它们需要依赖大量的有标签数据才能成功学习人工摘要的模式. 但是, 视频摘要任务需要在理解视频内容的前提下才能进行人工标注, 且具有一定的主观性, 这导致视频摘要标注困难, 且很难保证不同标注者人工摘要的一致性 [14,157,158]. 针对此问题, 研究人员在模型优化策略上已经作出了重大努

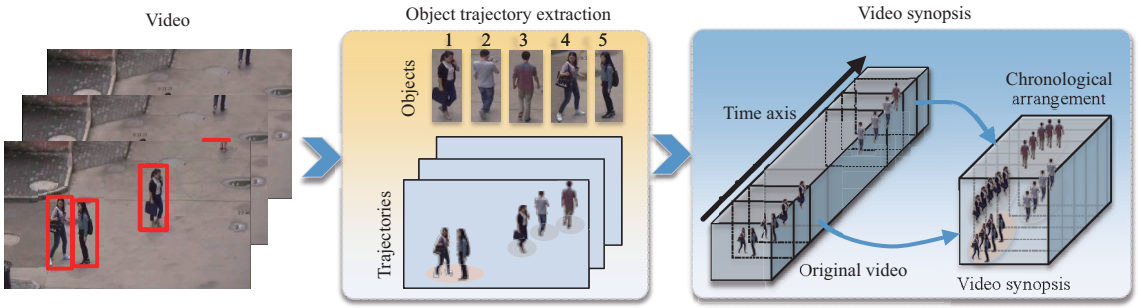

图 5 (网络版彩图) 视频浓缩示意图

Figure 5 (Color online) Demonstration of video synopsis

力, 比如: Mahasseni 等 [120] 提出视频摘要模型的对抗训练策略, 利用性质模型来反馈摘要生成器的优化过程. Zhao 等 [39] 设计了视频摘要的对偶学习策略, 利用视频重建过程指导摘要生成过程, 缓解了模型训练对标注数据的依赖. 但是, 完全摆脱人工标注, 实现视频摘要模型的高效无监督学习仍然任重道远.

(2) 视频摘要的在线实时生成问题. 目前的视频摘要方法大都是在对视频完整内容进行建模与分析的基础上, 进行关键帧和关键镜头的挑选, 这限制了视频摘要的在线生成能力. 在线字典学习的方法提供了视频摘要在线生成的新思路 [33, 65, 133], 但是该思想并没有在基于深度学习的方法中得到发展, 这与递归神经网络方法在建模视频全局信息的基础上进行摘要的思想相违背. 因此, 如何在保证质量的前提下实现视频摘要的在线生成, 对视频摘要的实用性至关重要.

(3) 视频摘要的评价问题. 实际应用中, 视频摘要需要满足语义理解层面的需求, 协助人们理解原视频的主要内容. 但是, 现有方法大都通过计算生成摘要与人工摘要的相似度进行算法性能评价. 这种评价方式忽略了每个关键帧和关键镜头的特异性, 以及视频摘要所需满足的语义需求, 导致其不能准确反映视频摘要的质量 [159]. 因此, 如何有效地对视频摘要的质量进行综合评价是亟待解决的关键问题.

4 视频浓缩

Rav-Acha 等 [160] 于 2006 年首次提出视频浓缩 (video synopsis) 的概念. 如图 5 所示, 该任务主要面向监控视频开展, 首先需要提取出视频中的运动目标, 然后对各个目标的运动轨迹进行分析, 再将不同的目标拼接到一个共同的背景中, 同时将它们以更为紧致的方式进行组合, 生成新的浓缩视频 [161, 162]. 总体来说, 视频浓缩任务的目的是在保持监控视频中目标运动信息的前提下, 尽可能地压缩数据量, 提高视频画面中的信息密度, 即通过以下方式提高单位数据量的信息提供能力:

$$\max_{O} \varphi = \frac{\sum_{o_i \in O} I\left(T_{o_i}\right)}{\sum_{o_i \in O} D\left(T_{o_i}\right)}, \tag{10}$$

其中, O 代表视频中目标的集合, T_{o_i} 代表目标 o_i 的运动轨迹.

在下面小节, 本文首先对比了视频中目标与轨迹的提取方法 (object and trajectory extraction), 然后针对单摄像头 (single-camera video synopsis) 和多摄像头 (multi-camera video synopsis) 的情况分类介绍了不同的视频浓缩方法.

4.1 视频目标与轨迹提取

目标提取是视频浓缩算法流程中的第一步. 运动检测是视频浓缩任务中常用的前景目标提取方法, 主要包括像素差分法 (pixel difference)、时间中值法 (temporal median)、运动向量法等 (motion vector) [160,163,164]. 这类方法较为简单, 在运动目标密集、拍摄场景复杂、背景信息变化较大的情况下稳定性较差. 为了解决这一问题, 研究人员提出了背景建模的算法进行前景目标的提取, 包括背景减除算法 (background cut)、层次化背景模型 (hierarchical background modeling) 和自适应背景模型等 (adaptive background modeling) [41,165,166], 这类算法对复杂场景变化具有较好的鲁棒性. 为了进一步提高目标提取的准确度, 基于目标检测与分割的方法也逐渐被提出, 如基于梯度方向直方图 (HOG) 和区域卷积神经网络 (R-CNN) 的目标检测方法 [42,43] 等. 虽然近年来目标提取方法已经取得了长足的进步, 但是由监控视频光照变化、目标遮挡等导致的目标丢失与误检问题仍然存在.

目标运动轨迹提取与分析是保证视频浓缩质量的关键步骤. 在检测到目标后, 现有方法利用目标跟踪算法将同一目标在相邻帧中的检测区域连接起来, 进而获取目标在视频中完整的运动轨迹, 常用的跟踪算法有基于卡尔曼 (Kalman) 滤波、粒子滤波、图匹配的方法等 [44,45,167]. 实际上, 目标跟踪的效果将直接影响视频浓缩的质量, 跟踪失败将导致目标轨迹中断、目标碰撞等问题, 进而影响后续浓缩视频的视觉效果和语义准确性. 为了提升视频浓缩的效果, 研究人员进一步根据目标类别、运动方向、动作类型等信息对目标的运动轨迹进行聚类 [168~171], 对相似的目标活动进行同时展示. 特别地, Lin 等 [172] 提出了基于局部区域学习的异常检测方法, 将异常的目标运动轨迹分离出来, 在视频浓缩中单独标记. Li 等 [19] 提出了复杂场景下群体关系保持的视频浓缩方法, 在目标轨迹提取的同时加入了群体判别模型. 目前来说, 目标轨迹的独立性和完整性仍然是研究人员的重点努力方向.

4.2 单摄像头视频浓缩

视频中目标的运动轨迹提取完成后, 需要采取空间换取时间的方式对目标运动轨迹进行重新排列, 提高空间利用率的同时缩减视频时长, 进而获得原视频的简短版本. 研究人员提出了多种多样的视频浓缩方法来尽可能地减少信息冗余, 同时保持目标在原视频中的时空关系, 且避免碰撞遮挡现象的发生. 例如, Rav-Acha 等 [160] 设计了全局能量函数, 对目标的活动、时间一致性和碰撞成本进行约束, 然后应用模拟退火方法进行能量函数优化. Pritch 等 [173] 在 Rav-Acha 等 [160] 的工作上进行了扩展, 首次提出了物体管道的概念, 用于描述目标的运动轨迹和活动信息, 对网络摄像头所拍摄的在线视频进行浓缩. 在该方法的启发下, 物体管道的概念在视频浓缩领域得到了广泛应用, 一直沿用至今. 在此基础上, 研究人员在提高浓缩率、减少目标碰撞现象, 以及提升算法效率方面取得了重大进展.

为了提高浓缩率, Correa 等 [174] 将同一物体在视频中的活动轨迹展现在一幅浓缩视频的画面中, 这样既展示了物体的运动过程又缩短了其在浓缩视频中占用的时间. Xu 等 [175] 将视频中的物体管道看作是一个集合, 并利用集合论的方法建立优化函数. 虽然这种方法能够获得更高的压缩率, 但是它破坏了目标的时序关系. 进一步地, Sun 等 [46] 对目标在浓缩视频中的时空位置进行优化. 在浓缩过程中, 活动目标不仅会被沿着时间轴移动, 它们的空间位置也会被改变. 该方法可以显著提高浓缩率, 但是浓缩视频中会出现行人在天空中行走等视觉歧义, 影响浓缩视频的可读性. 为了避免在空间优化过程中出现视觉歧义, Nie 等 [163] 提出了一种背景扩展方法, 使得背景能够适合位置改变后的物体, 如把视频中的马路拓宽以放下更多的车辆等. 这种方法在避免视觉歧义的同时提高了浓缩率, 但是只适用于画面中具有障碍物的视频, 如: 花坛、建筑物等. 总体来说, 提高空间利用率是保证浓缩率的主要方式.

为了避免或减少目标碰撞现象, 研究人员在视频浓缩优化过程中加入了各种各样的碰撞损失项. Hsia 等 [47] 在优化过程中只考虑了目标碰撞的能量函数, 通过计算每个目标管道之间的碰撞代价, 在时间维度重复移动管道, 直到生成无碰撞或可接受碰撞的浓缩视频. 与之不同的是, Li 等 [176] 提出自适应目标缩小的视频浓缩方法, 优化框架包括活动损失代价、碰撞损失代价和时空一致性代价. 该方法在发生碰撞现象时, 可以以目标的几何中心为原点, 对碰撞目标进行尺寸调节, 进而得到更加紧凑同时又包含更少碰撞遮挡现象的浓缩视频. 但是, 目标尺寸的变化会影响浓缩视频的视觉效果, 比如出现行人比汽车还大的情况, 对人们理解浓缩视频造成了一定的困扰. 为缓解此问题, Nie 等 [177] 将目标的尺寸和运动速度同时作为优化变量, 在出现碰撞遮挡现象时, 对目标的运动速度进行自适应调节, 以避免目标被过度缩小的问题. 进一步地, He 等 [45, 178] 对目标的碰撞状态进行了分类, 包含无碰撞、同一方向碰撞和相反方向碰撞 3 种类型. 在此基础上, 提出了目标碰撞的图优化模型, 提高浓缩密度的同时减少碰撞现象的产生.

为了提升视频浓缩的时间效率, 研究人员提出了基于运动像素的快速目标提取方法, 以及目标轨迹在线排列方法等. 其中, Yildiz 等 [179] 在轨迹分析时抛弃了基于目标提取的方法, 转而采用基于运动像素的方法, 通过动态规划在视频中寻找运动信息较少的时空区域, 同时利用正交投影得到可以丢弃的时空区域, 获得浓缩视频. 在此基础上, Vural 等 [180] 利用了人眼注视模型, 可以有效对监控视频中容易忽略的目标活动进行审查, 提高了浓缩视频的可靠性. 但是, 上述两种方法都是基于运动像素的方法进行目标提取, 往往导致目标不完整或者背景混杂的问题, 影响浓缩视频的视觉效果. 除了目标提取之外, 目标轨迹排列优化过程也占用了大量的时间. 模拟退火算法 (simulated annealing) 是视频浓缩任务中最常用的优化算法 [176, 181], 它面对复杂的优化问题可以跳出局部最优陷阱, 进而得到全局最优解. 但是, 模拟退火算法收敛速度较慢, 严重影响视频浓缩效率. 针对此问题, Ghatak 等 [182] 提出了 JAYA 算法 [183] 和模拟退火算法相结合的优化方法, 分别利用了两者的性能和速度优势, 在保证视频浓缩质量的同时, 提高了浓缩效率. Huang 等 [166] 强调了在线视频浓缩的重要性, 并提出了目标运动轨迹排列的在线优化算法, 实现在线目标排列的局部最优. 但是, 这种方法完全忽略了目标碰撞的问题, 而且需要人工设置阈值来确定浓缩效果. 为了提升在线视频浓缩的性能, Ruan 等 [184] 提出了动态图着色 (dynamic graph coloring) 模型, 对目标碰撞、轨迹和交互进行联合优化, 综合提高了在线视频浓缩的质量. 特别地, 研究人员意识到视频解码会增加视频浓缩算法的复杂度, 提出了在视频编码格式中直接进行浓缩的方法 [185, 186]. 但是, 视频编码格式中的目标提取较为困难, 导致这类方法在视频浓缩中的视觉效果难以保证.

4.3 多摄像头视频浓缩

由于监控摄像头的普及, 同一个场景中往往包含多个摄像头, 同一目标的活动也通常由不同的摄像头捕获. 针对分布式的视频监控网络, 多摄像头的视频浓缩具有更大的实用价值. 在此背景下, 研究人员逐渐意识到多摄像头视频浓缩任务的重要性. 该任务的关键问题是在浓缩之前进行多源视频的融合, 实现不同视角场景信息的对应. Zhu 等 [48] 在 2014 年首次提出了视野重叠的多摄像头视频浓缩方法, 它将所有摄像头的视野都转换到同一个平面中, 拼接获得全景视野, 并在此基础上将同一目标在不同摄像头中的轨迹与活动信息关联起来, 进而生成多摄像头的浓缩视频. 受此启发, Mahapatra 等 [171] 提出了类似的多摄像头全景拼接方法进行视频浓缩, 将所有摄像头的拍摄视野生成鸟瞰图, 利用各摄像头的目标运动信息分别进行视频浓缩. 此外, 该方法将视频中的活动分为步行、跑步、弯腰等类别, 可以生成特定类别的浓缩视频. 类似地, Zhang 等 [187] 提出了一种联合目标移动和多摄像头视角切换的视频浓缩方法, 并设计了由图割法和动态规划相结合的优化策略, 来对目标时序移动和视角切换进

行联合优化, 生成的视频浓缩简洁完整、通俗易懂.

以上方法都需要摄像头之间具有重叠视野, 进行全场景拼接. 针对摄像头之间没有视野重合的情形, Zhu 等 [49] 提出了基于目标重识别的多摄像头视频浓缩方法, 准确匹配同一目标在不同摄像头中的轨迹. 在优化过程中, 该方法设计了一种关键时间戳选择方法, 用于查找目标管道中的重要时刻, 比如: 出现、合并、分割和消失等, 然后设计了整个视野的全局能量函数来重新排列目标管道, 同时考虑了单一摄像机视角和全局视角的时间一致性. 虽然来自不同摄像头的目标轨迹在计算过程中被混合在一起, 但是浓缩视频展示时画面是分开的, 所以该方法本质上是利用多摄像头的信息来进行各个摄像头的视频浓缩. 与之不同的是, Hoshen 等 [43] 设计了摄像头的层次化结构, 在摄像头集群中定义了主摄像头和从属摄像头, 一旦在主摄像头中检测到目标活动, 则生成包含从属摄像头活动的浓缩视频, 为视频浓缩提供了更广阔的视角, 同时避免了目标重识别的不稳定性. 为了进一步拓展多摄像头的视频浓缩能力和效率, Lin 等 [50] 建立了视频浓缩的分布式处理框架. 该框架包含视频处理的各个步骤, 如初始化、背景建模、目标检测和跟踪、轨迹重排优化等, 并以并行方式进行计算. 分布式处理框架拓展了多摄像头视频浓缩的能力和效率, 是未来视频浓缩的重要发展方向.

4.4 视频浓缩评价

大多数视频浓缩方法都是无监督的方法, 在任意监控视频上即可进行实验. 视频浓缩需要考虑的因素较多, 浓缩视频的质量评价方式也较为复杂. 为了有效评价浓缩结果, 研究人员设计了多种多样的指标 [175, 176, 178, 185, 186]. 总体来说, 视频浓缩的评价指标具有一定的主观性, 主要围绕萃取信息的准确性和视觉效果而设计, 具体如下.

(1) 长度压缩率 (frame condensation ratio, FR) [176]. 计算浓缩后的视频长度与原视频长度的比值, 具体如下:

$$\mathrm{FR} = |\boldsymbol{S}|/|\boldsymbol{V}|, \tag{11}$$

其中, $|\boldsymbol{S}|$ 和 $|\boldsymbol{V}|$ 分别代表浓缩前后的视频帧数目. 可以看出, 压缩率 FR 越小说明得到的浓缩视频越简洁, 视频浓缩效果越好.

(2) 空间简洁率 (compact rate, CR) [175, 178]. 度量浓缩视频每一帧中前景目标占整个画面的平均比例, 定义如下:

$$\mathrm{CR} = \frac{1}{w \cdot h \cdot |\boldsymbol{S}|} \sum_{t=1}^{w \cdot h \cdot |\boldsymbol{S}|} \{1|\text{if } p\,(t) \in \text{foreground}\}, \tag{12}$$

其中, w, h 和 $|\boldsymbol{S}|$ 分别代表浓缩后视频帧的宽度、高度和数量, $p\,(t)$ 代表视频浓缩中的某一像素. 显然, CR 值越高说明浓缩视频越紧凑.

(3) 碰撞率 (overlap rate, OR) [178]. 统计浓缩视频中运动目标的碰撞情况, 即前景像素重合占浓缩视频总像素数的比例,

$$\mathrm{OR} = \frac{1}{w \cdot h \cdot |\boldsymbol{S}|} \sum_{t=1}^{w \cdot h \cdot |\boldsymbol{S}|} \{1|\text{if } p\,(t) \in \text{the collision foreground}\}. \tag{13}$$

碰撞率 OR 越高说明浓缩效果越差.

(4) 时序错乱率 (chronological disorder rate, CDR) [17, 186]. 统计浓缩视频中目标出现顺序错乱的情况占总目标数的比例,

$$\mathrm{CDR} = \frac{1}{m} \sum_{t=1}^{m} \left\{1|\text{if } o_t^V \neq o_t^S, \text{ where } o_t^V \in \boldsymbol{V},\ o_t^S \in \boldsymbol{S}\right\}, \tag{14}$$

表 5 视频浓缩方法分析

Table 5 Analysis on video synopsis approaches

Video synopsis	Type	Approach
Object trajectory extraction	Foreground extraction	Pixel difference [163], temporal median [160], motion vector [164]
	Background modeling	Background cut [41], hierarchical background modeling [166], adaptive background modeling [165]
	Object detection	HOG [43], R-CNN [42]
Single-camera video synopsis	Improving compression	Video narratives [174], set theoretical method [175], spatial-temporal optimization [46], background expanding [163]
	Reducing collision	Object collision loss [47], scaling-down objects [176], speed-size incorporating [177], graph model [45, 178]
	Improving efficiency	Pixel-based detection [179, 180], JAYA and simulated annealing [182], online optimization [166, 184], compressed domain [185, 186]
Multi-camera video synopsis	With overlap	Panoramic view [48], bird's eye view [171], view switch [187]
	Without overlap	Object re-identification [49], hierarchical camera structure [43], distributed processing [50]

其中, o_t^V 和 o_t^S 分别代表在原视频和浓缩视频中排序第 t 出现的运动目标, m 代表总目标数. 可以看出, 时序错乱率越高说明浓缩效果越差.

(5) 视觉效果 (visual effects) [19, 185]. 综合考虑目标混杂、背景缺损、轨迹断裂、视觉歧义等方面的缺陷, 对浓缩视频的视觉效果进行人工评价.

4.5 分析与讨论

视频浓缩任务从目标角度进行视频萃取. 如表 5 [41~43, 45~50, 160, 163~166, 171, 174~180, 182, 184~187] 所示, 视频浓缩方法在目标提取、轨迹分析、浓缩生成等方面取得了较大的进展. 视频浓缩呈现出多摄像头、跨场景和实时在线生成的发展趋势, 显著提高了监控视频中目标轨迹检索与分析的效率, 对公共安全、城市管理和司法取证等领域具有重要的应用价值.

虽然视频浓缩已经取得了重大发展, 但是仍然存在一些关键问题亟待解决.

(1) 视频中目标与轨迹提取的基础问题. 视频浓缩是凌驾于目标检测与跟踪任务之上的高层语义任务, 浓缩质量严重依赖目标与轨迹的提取结果, 但是由于视频画面中的光照变化、场景转换等问题, 以及目标之间的遮挡、运动、交互等问题, 现有视频浓缩任务中利用的目标检测、背景提取、轨迹跟踪和目标再识别等方法很难准确无误地实现目标与轨迹的提取. 实际上, 即使利用现有性能最优的目标检测器与跟踪器, 也无法避免目标缺失和轨迹破坏的问题 [188~190]. 然而, 公共安全、刑事侦查等应用场景对目标和轨迹的完整性要求较高. 该问题若不能得到很好的解决, 将严重影响视频浓缩在相关场景中的实用性.

(2) 浓缩视频的全天候生成问题. 监控摄像头具有全天候拍摄的特点, 但是现有的视频浓缩方法大都针对日间场景的监控视频进行处理, 缺乏对夜间视频的浓缩能力. 夜间视频浓缩的特点主要在于低光照条件下目标与轨迹的提取更为困难, 以及红外夜视摄像头的视频画面相对于可见光摄像头存在显

A man is playing the basketball game.

图 6 (网络版彩图) 视频描述示意图

Figure 6 (Color online) Demonstration of video captioning

著差异. 实际上, 夜间目标相对较少, 可以实现更高的浓缩率, 而且夜间也是异常事件突发的时间. 因此, 非常有必要结合夜间视频的特点, 对夜间视频浓缩进行研究.

(3) 浓缩视频的端到端生成问题. 由于视频浓缩任务的复杂性, 现有方法通常沿用目标提取、轨迹分析、浓缩视频生成的流程框架, 这往往会导致计算误差不断累计, 例如: 目标提取不准确会导致轨迹不完整, 最终影响视频浓缩效果. 然而, 端到端学习的方法可以实现输入数据和输出结果的连接, 在不同的视频分析任务中取得了优异的效果 [51, 191, 192]. 因此, 如何实现浓缩视频的端到端生成是该任务急需解决的关键问题.

5 视频描述

如图 6 所示, 视频描述是将计算机视觉和自然语言处理研究连接起来的跨模态视频萃取任务, 在对视频进行语义理解的基础上, 生成文本语句描述视频内容, 从语义角度进行视频萃取, 对基于内容的视频检索、人机交互和盲人导航等具有广泛的应用价值. 总体来说, 视频描述是通过将视频数据转化为文本数据, 视觉信息转换为语义信息的方式进行视频萃取. 由于描述语句可以最大可能地保持原视频的语义信息, 而文本的数据量显著小于原视频的数据量, 显然视频描述生成后单位数据量的信息提供能力

$$\varphi = \frac{I(c)}{D(c)} \tag{15}$$

被大大增强, 其中, c 代表所生成的描述语句.

视频描述具有悠久的发展历史, 主要分为基于语句模板的传统方法 (template based video captioning) 和基于 CNN-RNN 框架的深度学习方法两个发展阶段. 接下来, 本节将分别进行介绍.

5.1 基于语句模板的视频描述

基于语句模板的视频描述方法分为两步 [193~195]. 第 1 步, 为生成语句中的每个成分训练单独的分类器, 包含主语、谓语、宾语等. 第 2 步, 依靠语句模板, 利用概率图模型等方法将各个成分融合为一个完整的句子.

针对第 1 步, 研究人员提出了许多方法来检测视频中的元素, 如物体、人、动作和事件等. 由于语句模板是视频描述较为早期的方法, 这类方法所采用的目标检测算法也大多是传统方法, 如基于边缘检测或颜色匹配的方法 [196]、基于方向梯度直方图 (HOG) 的方法 [20]、基于尺度不变特征变换 (SIFT) 的方法 [22] 和基于局部可形变模型 (DPM) 的方法 [52] 等. 动作和事件检测方法通常借助基于时空特征点的光流方向直方图特征, 利用动态贝叶斯网络 (dynamic Bayesian network, DBN) [197] 和隐马尔可夫模型 (hidden Markov model, HMM) [198] 等进行检测与识别. 上述方法将视频中的视觉元素进行

单独检测, 当元素较多时, 无法确定用于描述的目标和活动. 为解决该问题, 研究人员提出了利用各元素之间潜在的语义关系, 对涉及多人、多物、多行为的事件进行联合建模, 进而对视觉特征进行综合提取 [53, 54].

针对第 2 步, 研究人员设计了各种各样的语句模板. 语句模板通常由词汇表、语法和模板规则 3 部分组成. 其中, 词汇表是用于生成视频描述的词汇集合. 语法代表基本的语言规律, 用于保证生成描述语句在语法上的正确性. 模板规则由用户定义, 用于指导挑选合适的词汇来生成句子. 为了准确描述视频中的活动, 常用的语句模板结构如下:

主语(subject)+ 谓语 (verb)+ 宾语 (object)+ 状语 (adverbial)

例句: A boy is playing basketball on the ground,

其中, 下划线部分表示可选项. 利用语句模板, Kojima 等 [196] 提出了一个简单的视频描述方法, 它只关注于包含单一人物和单一活动的简单视频, 通过检测人物的头部和手臂确定视频中的活动, 然后依据语句模板生成视频描述. 该方法是语句模板在视频描述任务中的成功尝试, 但是它无法应用于目标和活动更为复杂的场景中. 为解决该问题, Hanckmann 等 [199] 提出了为多个人物的多个活动生成视频描述的方法. 该方法同时考虑了人与人、人与物之间的交互, 利用特征包 (bag-of-features) 方法作为动作检测器, 对视频中发生的不同动作进行定位与分类, 并将其与对应的人物关联起来, 用于生成多人物和多活动的视频描述. 虽然该方法相较于 Kojima 等 [196] 的方法取得了较大的进步, 但是仍然只能描述 48 个活动类别, 难以应用到实际开放场景中. Krishnamoorthy 等 [195] 提出了早期的实际开放场景的视频描述方法, 它利用网络词汇语料库作为辅助信息, 对不同类别的 YouTube 视频进行描述生成, 是基于语句模板匹配的视频描述方法在实际场景中的成功尝试. 在此基础上, 研究人员提出了多种方法来提高语句模板算法的性能, 包括利用概率图模型、目标检测模型、语言统计模型来扩充词汇表, 增加对目标和活动的描述类型, 提高对视觉和文本信息的建模能力 [193, 194]. 但是, 语句模板太过固定, 灵活性不足, 难以对视频中丰富的语义信息进行建模, 导致生成的文本描述句式单一、质量不高. 相比之下, 基于数据驱动的深度学习方法更适合对视频中丰富的语义信息和文本中复杂的语言规律进行建模. 因此, 随着深度学习的发展, 基于模板匹配的方法逐渐被基于 CNN-RNN 框架的方法所取代.

5.2 基于 CNN-RNN 框架的视频描述

近年来, 随着卷积神经网络 (CNN) 和递归神经网络 (RNN) 的发展, 基于 CNN-RNN 框架的视频描述算法逐渐被提出 [55, 200, 201]. CNN-RNN 方法以 CNN 作为视觉信息编码器, RNN 作为描述文本生成器, 利用损失函数度量生成描述和人工描述之间的相似度, 通过有监督的训练获得有效的视频描述模型 [202, 203]. 此类方法的研究重点在于视觉特征编码器和描述文本生成器的设计上 [51, 204, 205], 具体如下.

5.2.1 视觉特征编码器设计

早期的方法中, 视频描述的生成直接由图像描述方法扩展而来. 借鉴图像描述方法, 视频的视觉特征来自于每一帧 CNN 特征的均值池化, 然后将池化后的特征向量输入到 LSTM 中逐步生成视频描述 [55, 206]. 这类方法注重空间特征的提取, 却忽略了视频帧的时序信息, 降低了视觉特征的判别性, 造成了一定的信息损失. 为了弥补这一缺陷, S2VT 方法 [24] 中 LSTM 同样用于视觉特征编码. 这种方法包含两个 LSTM, 其中第 1 个 LSTM 用于对 CNN 提取的视频帧特征进行编码, 第 2 个 LSTM 利用编

码后的视觉信息和文本信息实现视频描述的逐步生成. 通过利用 CNN 和 LSTM 对视频帧进行编码, 可以同时挖掘视频的空间和时间信息, 进而显著提高了视频描述的质量. 但是, 考虑到 LSTM 在长度为 100 以内的视频帧序列中才能取得最好的表现, 而需要描述的视频往往具有数百帧, 传统的 LSTM 只能通过对视频帧进行降采样来处理这种视频帧序列, 这会造成不可避免的信息损失. 基于此, Pan 等 [25] 提出层次化的 LSTM, 其包含两层网络, 每一层都是一个较短的 LSTM. 该方法首先将视频帧序列分割为短的视频片段, 并将分割后的片段输入到第 1 层 LSTM, 其最后一步的隐状态作为该视频片段的编码特征输入到第 2 层, 而第 2 层 LSTM 最后一步的隐状态作为整个视频的编码特征输入到描述文本生成器. 进一步地, Baraldi 等 [201] 在层次化 LSTM 的基础上考虑视频结构信息进行视觉特征编码, 其中第 1 层进行视频镜头边界检测并获得镜头特征编码, 第 2 层用于视频全局特征编码, 进一步提高了视觉特征的判别能力. 此外, 为了更好地从视频中获得关于描述主体的有效特征, Zhao 等 [56] 在视频浓缩任务的启发下, 设计了物体的管状特征 (tube feature). 该方法首先使用 Faster-RCNN 来逐帧检测视频中的目标. 然后, 将检测到的目标区域按照空间位置关系依次连接形成物体管道, 进而获得了物体在视频中的运动轨迹. 最后, 该方法利用 LSTM 对物体管道进行时序编码获取管状特征, 减少无关背景信息的干扰, 提高了视觉特征的判别能力. 但是, 管状特征的提取依赖检测器的性能, 而且无法进行端对端的模型学习, 这在一定程度上增加了视频描述的不稳定性和累计误差.

为了实现视觉特征的自适应提取, 科研人员在大脑注意力机制的启发下, 将注意力模型引入到视觉信息编码器中 [51,57,207]. Yao 等 [58] 首先在视频描述中采用了视觉注意力模型, 该方法假设每一个视频帧对于生成描述具有不同的相关性, 在视觉信息编码时, 应该强调相关的视频帧而减少无关视频帧的影响. 基于此, Yao 等采用了 2D CNN 和 3D CNN 的双流特征框架, 同时挖掘视频时空信息, 并为每一个输入的视频帧学习自适应的权重, 同视频帧的 CNN 特征一起用于视觉信息的编码. 在此基础上, Cherian 等 [202] 设计了时空注意力模型来对视觉信息进行时空联合编码. 类似地, Li 等 [59] 设计了层次化的视觉注意力模型, 可以自动聚焦于与描述文本最相关的视频帧以及帧内最显著的区域中, 实现了视觉特征的层级编码, 减少了无关视频帧和背景信息的干扰, 提高了视觉特征的判别能力.

为了实现视觉特征的精准提取, 语义信息被用于指导视觉特征的提取过程. Pan 等 [200] 将视频的属性标签, 如: 女孩、马、蓝天等, 作为额外的语义信息生成视频描述. 该方法设计了一个多方面的注意力机制模型, 同时为视频帧和语义属性分配编码权重, 对视觉信息和语义信息进行融合. 实际上, Yu 等 [208] 所提出的语义注意力模型也能达到类似的效果, 而且不需要额外的语义属性标签. 另外, Venugopalan 等 [209] 通过挖掘更有效的文本特征提高视频描述的精度. 受此启发, Zhao 等 [210] 意识到跨模态信息融合对视频描述任务的重要性, 提出了视觉和文本特征的联合注意力模型, 联合利用分层视觉注意力模型、词组文本注意力模型和跨模态特征协调器, 实现视觉和文本特征的自适应编码与跨模态融合, 显著提升了视频描述的质量.

5.2.2 描述文本生成器设计

大多数视频描述生成器都是在 LSTM 的基础上建立的 [12,24,211,212]. 在生成过程中, LSTM 以视觉特征和上一步生成的文本特征作为模型输入, 实现描述文本的逐字生成. 为了提高生成描述的丰富程度以及与人类表达的相似性, 基于强化学习和生成对抗网络的方法也逐渐被提出 [213,214]. Ren 等 [215] 采取一种迭代训练强化学习的方法, 将生成描述的错误信息输入到模型中进行迭代优化, 直到得到令人满意的结果. 为了对视频中的精细活动进行准确描述, Wang 等 [216] 提出了层次化的强化学习策略, 该方法将完整的视频描述分成小段, 并建立分层序列模型对其进行建模, 其中低层模型用于词汇逐字生成, 高层模型在语义理解的基础上进行文本片段拼接, 进而达到描述文本精细化生成的目的.

表 6 视频描述数据集统计

Table 6 Statistics on video captioning datasets

Dataset	Year	Size	Average duration (s)	Sentences	Average sentences	Vocabulary size	Type
MSVD [220]	2011	1970	10	70028	35.55	13010	Open domain
TACoS [221]	2013	7206	360	18227	2.53	28292	Cooking
M-VAD [222]	2015	48986	4	55904	1.14	17609	Movie
MPII-MD [223]	2015	68337	6	68375	1.00	24549	Movie
MSR-VTT [224]	2016	10000	20	20000	2.00	29316	Open domain
Charades [225]	2016	9848	30	27847	2.83	–	Human activity
VTW [156]	2016	18100	90	44613	2.46	–	Open domain

为了对描述生成器提供更多的指导信息, Li 等 [60] 提出多任务的强化学习策略, 在视频描述任务之外开发标签学习任务, 利用人工描述中的实词作为视频标签, 在训练描述生成器的同时, 通过强化学习策略约束模型对视频标签的预测能力, 从而为描述生成器提供更多的监督信息, 以缓解模型过拟合问题. 类似地, Zhang 等 [217] 在描述生成器的基础上增加了视频信息重建器, 对生成描述进行建模, 并将其文本信息与视频编码信息进行对比, 约束生成描述对原视频内容的重建能力. 总而言之, 强化学习的优化策略是提高模型训练性能的一种重要手段.

为了进一步提高视频描述的准确性, Dai 等 [61] 利用生成判别模型来生成视频描述, 在描述生成器的基础上添加文本判别器, 用于判断生成描述语法和语义上的准确性. 在判别器的作用下, 该算法生成的视频描述具有更少的语法和语义错误, 且更符合人类的表达方式. Song 等 [218] 提出了基于多模态随机递归神经网络的描述生成模型, 利用随机变量对视频数据中的不确定性进行建模, 在不同随机因素的情况下生成多个句子来描述视频内容. 类似地, Nabati 等 [203] 设计了并行级联递归神经网络结构作为描述生成器, 可以同时提高视频描述的质量和生成效率. Rahman 等 [62] 提出了视频密集描述生成策略, 联合利用视觉、听觉和文本信息构建跨模态注意力模型, 同时进行时序活动检测与描述生成, 为视频中的各项活动单独生成描述文本. 特别地, Wu 等 [219] 抛弃了基于 LSTM 的描述生成器框架, 创造性地设计了基于层次化记忆力模块 (hierarchy memory decoder) 的描述生成器, 这种解码器缓解了 LSTM 长时信息遗忘的问题, 而且层次化结构有助于捕获视频帧和单词的时序依赖关系, 是解码器结构设计的一次成功尝试. 随着视频描述任务的发展, 传统 LSTM 结构中存在的问题愈发暴露, 在 LSTM 的基础上进行改进以解决实际问题, 以及设计新的网络结构是目前描述生成器的重点研究内容.

5.3 视频描述数据集与评价指标

视频描述任务常用的公开数据集如表 6 [156, 220~225] 所示, 部分样例在图 7 中展示. 从表中可以看出, 现有的视频描述数据集大都在 2011~2016 年建立, 包含烹饪、电影、人物活动和开放的视频类别. 数据集尺度较大, 通常包含成千上万个视频片段, 每个视频片段拥有一个或多个人工生成的视频描述. 其中, MSVD 数据集中每个片段的平均描述语句数量超过 35 个, 这对训练描述生成器的灵活性, 以及学习人类语言的表达方式至关重要. 每个数据集包含的词汇表大约在 1 万 ~ 3 万之间, 研究人员在设计算法时往往会剔除词汇表中的生僻词, 以此减少词汇表中单词的数量, 降低学习难度 [24, 57, 59].

视频描述任务常用 4 个评价指标来度量生成描述的质量, 它们分别是 BLEU [226], ROUGE-L [227], CIDEr [228] 和 METEOR [229]. 其中, BLEU 包含 4 个版本, 即 BLEU 1~4. 实际上, 上述评价指标都

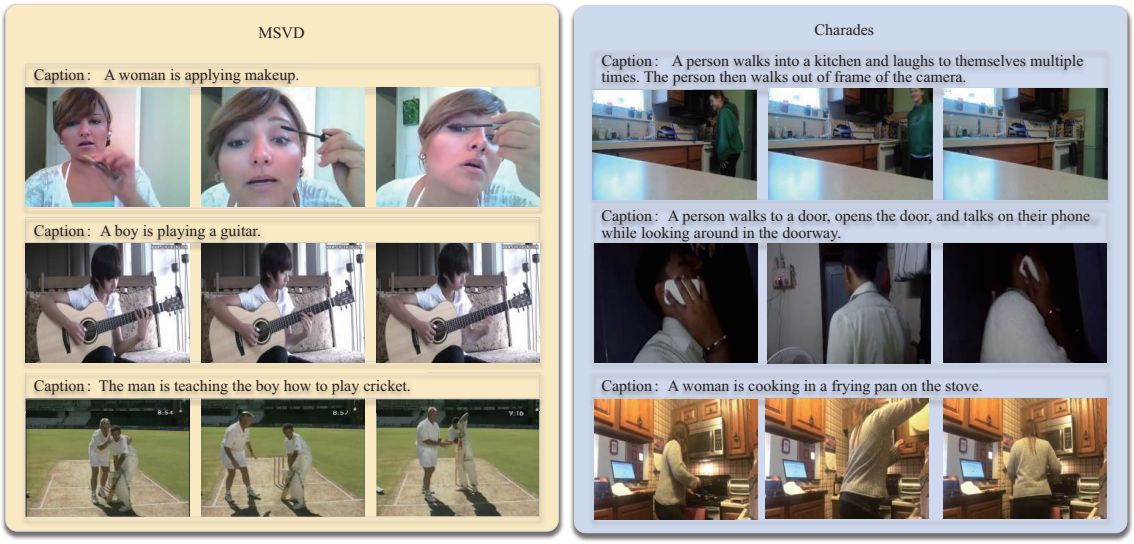

图 7　(网络版彩图) 视频描述数据样例展示

Figure 7　(Color online) Examples of video captions

来自于机器翻译任务. 由于视频描述和机器翻译的评价都是对比人工生成语句和机器预测语句的相似性, 因此它们也被广泛应用于视频描述任务中. 需要注意的是, 上述指标的分数与视频描述的质量是正相关的, 也就是说分数越高代表生成的描述文本越准确. 此外, 由于上述定量评价指标的局限性, 人工评价也经常用于度量生成描述的语法准确性、语义完整性和表达灵活性等, 实现对生成描述质量的综合评价 [210].

5.4　分析与讨论

视频描述任务从语义角度进行视频萃取. 如图 8 [18, 24, 25, 58, 194, 201, 206, 210, 216, 230~241] 所示, 近年来伴随着深度学习的快速发展, 视频描述任务完成了从基于语句模板匹配到基于 CNN-RNN 学习框架的跨越, 在大规模数据集的驱动下, 显著提高了生成描述的质量. 目前来看, 在 CNN-RNN 框架的基础上进行注意力模型、时空联合模型、双流特征结构等视觉特征提取器的改进, 以及强化学习策略、生成对抗网络、层次化结构等描述生成器的改进, 是现有方法的主流趋势和突出贡献. 此外, 研究人员也逐渐意识到现有框架的不足, 开始依据视频描述任务的特点, 进行全新的框架设计, 提高了视频描述任务的信息萃取能力.

虽然近年来视频描述已经取得了重大发展, 但是仍然存在一些关键问题亟待解决.

(1) 视频描述文本与视觉元素的对应性不强. 现有方法只是将视觉信息转换为文本形式进行表达, 忽略了视觉和文本的对应, 即无法利用文本信息自动找到视频画面中相对应的视觉元素. 实际上, 视觉信息和文本信息的对应可以实现视频内容的结构化解析, 提升语义理解水平. 另外, 利用文本语句进行相关视觉元素的检索, 可以促进其他视频分析任务的发展, 如视频摘要和浓缩等. 因此, 加强描述文本与视觉元素的对应性是提高视频描述信息萃取能力的关键问题.

(2) 文本信息利用不充分. 视觉和语言是人类认知和理解世界的两种最重要的方式, 视频描述是利用计算机视觉和自然语言处理技术将它们桥接在一起的关键任务之一. 但是, 现有方法将研究重点放在了视觉特征编码上, 忽略了文本信息的重要性. 虽然有些方法在传统文本特征的基础上进行了改进, 如文本注意力模型、语义标签建模等, 但仍然不足以实现文本信息的精准编码, 以及视觉和文本信

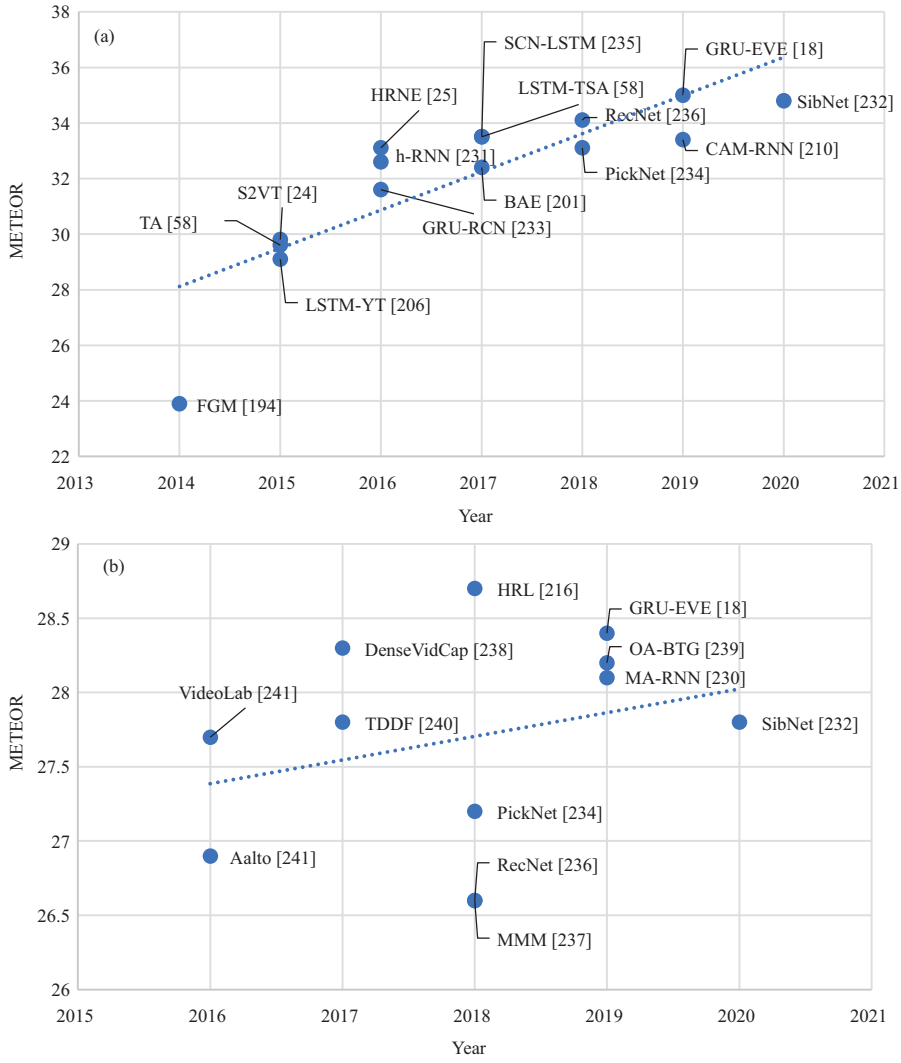

图 8 (网络版彩图) 视频描述结果在 (a) MSVD 和 (b) MSR-VTT 数据集上随年份变化曲线

Figure 8 (Color online) Video caption results vary by year on (a) MSVD and (b) MSR-VTT datasets

息的有效融合. 这在一定情况下限制了视频描述的质量.

(3) 视频描述的评价方式不够完善. 考虑到机器翻译和视频描述任务在输出结果上的相似性, 现有的视频描述质量评价指标大都是从机器翻译任务中直接迁移过来的, 相应地也继承了机器翻译中评价指标存在的问题. 实际上, 现有评价指标都只注重于人工标注与生成语句在文字上的相似度, 无法度量两个句子的语义相似度. 然而, 视频描述任务具有很强的语义需求, 文字相似并不能保证语义相似. 因此, 视频描述质量的评价仍然是需要重点攻克的学术难题.

6 视频萃取面临的挑战与未来发展趋势

研究人员依托视频摘要、浓缩和描述任务分别从内容、目标和语义角度进行视频萃取. 近年来, 随着图像分类、目标检测、轨迹跟踪、文本生成等基础问题在深度学习的助力下实现了跨越式发展, 视

频萃取技术也有了较大的进步. 在视频数据爆炸式增长的背景下, 视频萃取的需求日益增加, 已成为人工智能领域的热门研究课题. 本节首先分析了视频萃取面临的诸多挑战, 然后对未来的发展趋势进行了展望, 最后提出了一些值得思考的开放性问题.

6.1 视频萃取面临的挑战

现在视频萃取研究仍然处于初级发展阶段, 存在理论研究不充分、应用探索不系统等问题. 近年来, 视频获取方式的快速革新和拍摄需求的多样化发展, 对视频萃取带来了新的挑战. 回顾文章开头处关于数据、信息和知识的两个最基本的公式:

$$\Psi = \frac{\boldsymbol{I}}{\boldsymbol{D}}, \tag{16}$$

$$\kappa = \langle \boldsymbol{I}, \Lambda \rangle = \langle \Psi \odot \boldsymbol{D}, \Lambda \rangle . \tag{17}$$

考虑到视频数据、信息与人的认知的多维性, 上式中信容 Ψ, 数据量 $\boldsymbol{D}$, 信息量 $\boldsymbol{I}$ 和人的认知能力 Λ 都拓展为向量的形式. 简要来说: (1) 视频信息量 $\boldsymbol{I}$ 与数据量 $\boldsymbol{D}$ 的比值代表视频的信息提供能力, 视频萃取在保持信息量的前提下减少数据量, 进而提升视频的信容 Ψ. (2) 人观看视频所获得的知识量是信息量 $\boldsymbol{I}$ 在认知能力 Λ 上的投影, 所以视频萃取的形式要考虑人的认知习惯. 视频萃取任务面临的挑战围绕上式中的变量展开, 包括数据量 $\boldsymbol{D}$, 信息量 $\boldsymbol{I}$ 和人的认知能力 Λ, 主要体现在数据建模、信息处理和个体认知 3 个方面, 具体如下.

6.1.1 视频数据建模

随着拍摄设备的升级, 视频数据呈现出高维度、高冗余和高混杂的特点. 首先, 为了达到更好的视觉效果, 视频的空间分辨率越来越高. 普通手持设备, 如: 智能手机、卡片机、GoPro 等, 已经具备了 720 p, 1080 p, 2 K, 4 K 的拍摄能力. 然而, 现有流行的视频空间表征方法大都习惯于处理相对较低分辨率的视频, 尤其对于卷积神经网络, 低分辨率是对效率和性能妥协的结果. 对高分辨率图像进行空间降采样预处理, 然后利用传统方法进行特征提取, 是现在常用的图像空间建模方法. 这显然会丢失大部分细节信息, 无法适应视频的高分辨率发展趋势. 因此, 如何对高分辨率图像和视频进行建模是对目前空间表征模型的巨大挑战.

另外, 为了捕获更为流畅和细致的运动信息, 高质量视频的帧率也越来越高, 不同于传统 24 fps (frames per second), 25 fps 和 30 fps 的视频, 近年来 60 fps 以及 120 fps 的视频也越来越多. 在这种情况下, 若干秒的短视频就会有成千上万个视频帧. 然而, 目前仍没有针对超长视频序列非常有效的时序建模方法. 实际上, 对于最擅长处理长序列的 LSTM 模型, 也仅能在序列长度小于 100 的视频上取得较好的结果 [24]. 这使得目前的视频萃取方法只能处理较短的视频, 即使视频摘要这种专门处理长视频的任务, 视频长度通常也小于 10 min. 因此, 视频的高帧率发展趋势, 以及长视频的萃取需求, 极大地增加了时序编码的难度.

此外, 随着拍摄设备的普及, 多设备共同拍摄的情况越来越多, 如街角的多个监控摄像头等. 为适应不同的环境, 光谱相机也逐渐普及, 如红外相机等. 如何实现多视角数据的联合表征和多源数据的有效融合, 进而削减多源数据的混杂性, 以及在此基础上保证后续任务的性能和时间效率, 也是视频萃取的一个重要挑战.

6.1.2 视频信息处理

视频数据所携带的信息具有多维和多模态特性. 首先, 视频中的信息是多维的, 包含目标、场景、

运动、语义等几个方面, 这也是本文倾向于用向量元素表示式 (16) 和 (17) 的原因. 现有的视频萃取方法往往针对特定方面信息进行萃取, 例如视频摘要方法重点关注场景和目标信息的萃取, 视频浓缩方法致力于提取目标和运动信息, 视频描述方法聚焦于语义信息的编码. 然而, 视频的多维信息并不是独立显式存在的, 比如目标和场景信息同时存在于画面中, 运动信息记录了目标信息的时序变化, 语义信息需要借助类别标签或者自然语言处理任务的协助才能表现出来. 实际上, 如何对视频中的多维信息进行精准的提取是非常困难的, 这也是不同视频萃取任务存在的共性基础问题.

另外, 视频的图像、音频和字幕中蕴含了丰富的视觉、听觉和文本信息, 在不同模态信息的共同作用下, 视频才能为观众提供直观的视听感受. 现有的视频萃取技术大多只利用了视频的单一模态信息, 其中视频摘要和浓缩都只利用了视觉信息, 视频描述将视觉信息转换为文本信息. 3 个任务没有实现多模态信息的联合利用与萃取. 实际上, 多模态信息具有一定的互补性, 在视频萃取中融入多模态信息, 以及实现多模态信息的联合萃取, 将显著提高视频萃取的应用价值 [218, 230, 242]. 因此, 如何联合利用视频中的图像、音频和字幕等信息载体, 探索多模态信息融合机制, 是视频萃取的重大挑战.

6.1.3 个体认知探索

视频萃取的目的之一是让人更加高效快速地从视频中获取有效的信息, 所以视频萃取是需要以人的认知能力为导向的. 现有方法都采取符合人类普遍认知习惯的方式进行视频萃取, 如视频摘要和浓缩都从视觉角度进行萃取, 视频描述从文本角度进行萃取. 但是, 实验表明, 不同的人对视频内容的关注点不同, 人工萃取的信息也就不同, 尤其是对于视频摘要和视频描述任务, 从各大数据集的标注结果可以看出, 不同人对于同一视频生成的摘要和描述千差万别 [155, 220]. 这说明个体的认知习惯具有较大的差异性, 相应的视频萃取也需要具有个性化的特点. 现有方法基于统计理论的模型学习可以在一定程度上进行信息精准萃取, 但是无法满足不同人的个性化需求.

另外, 缺乏对个体认知的探索, 也导致了视频萃取后信息质量难以评价的问题, 统计相似度并不能准确反映不同个体对萃取信息的满意程度 [159]. 这也导致视频萃取缺乏明确具体的优化方向, 以至于视频萃取模型缺乏有效的指导信息, 进而影响模型的优化过程. 因此, 对个体认知能力进行探索, 进而将个体差异引入到视频萃取的各项任务中, 是目前视频萃取面临的重大挑战, 也是迫切需要解决的关键问题.

6.2 未来发展趋势

面对视频数据的爆炸式喷涌, 以及人们日益增长的视频智能分析需求, 人工智能领域迫切需要视频萃取的进步与发展. 针对目前所面临的主要挑战, 本文对视频萃取研究的未来发展趋势进行了预测, 具体包含深入基础问题研究、扩展数据处理方式、开发新的信息萃取应用 3 个方向.

6.2.1 深入基础问题研究

虽然随着人工智能的发展, 尤其是深度学习技术的革新, 视频萃取取得了重大进步, 但是仍然存在许多基础问题尚未解决.

首先, 如前文所述, 视频萃取是一个提取、浓缩和跨模态转换视频信息的过程. 那么, 研究人员非常有必要对视频信息进行度量, 即原视频的信息量、萃取后的信息量, 以及在萃取过程中丢失的信息量. 如何对视频信息量进行度量对视频萃取具有重要的理论意义.

其次, 视频萃取的一个重要功能是帮助观看者快速理解视频内容, 但是不同的观看者从同一视频中接收到的信息是不同的, 如何建模观看者和视频之间的信息交互是实现精准信息萃取的关键, 也是

个性化信息萃取的必经之路, 对视频萃取至关重要.

此外, 深度神经网络虽然具有强大的学习能力, 但是存在可解释性差的问题, 部分情况下模型性能的提高并不具有理论价值, 难以对视频萃取的发展提供良性指导. 深度学习的 “黑盒子” 性质导致模型的可靠性不可度量, 严重限制了视频萃取在国家重大战略需求上的应用, 包括国防建设、航空航天、精密仪器、装备制造等领域.

实际上, 视频萃取中尚未解决的基础问题还有很多, 相信未来针对视频萃取中的关键基础问题展开研究, 从理论上度量视频信息量、建模信息交互, 以及提高深度学习的可解释性, 将会是视频萃取发展的重要趋势.

6.2.2 扩展数据处理方式

现阶段的视频萃取方法, 大都针对用户视频 (视频摘要和描述) 和监控视频 (视频浓缩) 中的视觉信息展开研究. 然而, 视频数据具有多样性. 首先, 视频内部具有图像、文本和音频的多模态数据. 任一信息模态的缺失, 都会对观众理解视频内容造成不同程度的影响, 进而降低观看体验. 因此, 单一模态信息的处理限制了视频萃取的质量和应用推广. 为了在保留视频多模态特性的基础上进行更综合的信息萃取, 给用户身临其境的感受, 相信视频多模态数据的联合表征与协同融合将会是视频萃取后续发展的重要研究方向.

进一步地, 视频的类别也具有多样性, 根据主题可以分为体育、新闻、旅游、综艺、记录、广告、电视剧、电影、监控等. 现有视频萃取方法大都利用通用模型进行数据建模, 如卷积神经网络进行空间建模, 递归神经网络进行时序建模. 然而, 不同种类视频数据包含信息的侧重点不同, 比如监控视频的局部目标和运动信息较为重要, 旅游视频的全局场景画面信息更为重要. 因此, 数据表征是视频萃取的基础, 有必要对不同种类视频的数据建模进行细化处理, 这也是为达到更好萃取效果的必然趋势.

随着技术的革新, 除了自然光下拍摄的 RGB 视频外, 红外、遥感等技术也逐渐用于视频拍摄过程, 比如红外监控摄像头可以在夜间观察到清晰的目标, 高分卫星遥感可以在高空拍摄城市全天候的动态视频. 在不同传感器的帮助下, 视频拍摄的手段和获取的信息越来越多. 融合多源数据进行联合视频萃取, 将会是一个新的发展趋势.

6.2.3 开发新的视频萃取应用

随着基础问题研究的推进, 以及视频数据处理方式的扩展, 视频萃取新的应用方向亟待开发, 主要分为以下 3 个方面.

(1) 结合人类的认知习惯, 探索视频萃取的新形式. 人工智能的发展对视频智能分析的需求越来越大, 仅仅利用视频摘要、浓缩、描述的单一任务进行视频萃取难以满足人们的多样化需求. 实际上, 综合利用不同的视频萃取任务, 开发视频萃取的新形式, 可以提高萃取信息的精准度、多样性和趣味性. 目前已经有科研人员开始研究多任务的视频萃取框架, 如: 联合利用视频摘要和描述任务自动生成视频标题 [156], 联合利用视频浓缩和描述任务进行视频结构化语义理解 [243] 等. 为满足人们的多样化需求, 在视频摘要、浓缩、描述任务上进行新的探索, 以及开发多任务联合的视频萃取方式, 将会是提高视频萃取应用价值的重点研究内容.

(2) 视频萃取的实时性研究, 与移动平台的应用推广. 近年来, 智能手机、GoPro、微型无人机, 以及可穿戴摄影设备逐渐普及, 它们成为视频数据获取的主要途径. 为了提升视频质量、实现自动剪辑、缓解存储压力、方便视频管理, 在移动端直接进行在线视频萃取的需求越来越大. 但是, 现有方法大都在视频全局信息理解的基础上进行建模, 无法实现信息的实时萃取. 鉴于此, 未来工作中, 探索新的视

频数据建模方式、提升算法效率、降低硬件依赖、减轻模型大小, 并在此基础上开发移动端在线视频萃取的能力, 将会是未来视频萃取的重要发展趋势之一.

(3) 应用到其他任务中, 提高视频智能分析的性能和效率. 随着人工智能的发展, 基于视频数据的智能分析任务越来越多, 如人机交互、视觉导航、异常检测、辅助驾驶、自动剪辑等, 视频智能分析的需求呈现出多样化发展的趋势. 实际上, 这些任务都严重依赖视频萃取的效果, 例如: 人机交互依赖视频描述任务中的语义信息萃取, 异常检测依赖视频浓缩任务中的目标信息萃取, 智能剪辑依赖视频摘要任务中的内容信息萃取等. 因此, 通过视频萃取任务获取更为简洁的视频信息, 提高视频智能分析的效率和性能, 将会是未来视频萃取的研究热点.

6.3 开放性问题讨论

(1) 视频的信息如何度量? 信息论中引入熵 (entropy) 的概念对随机变量 X 的信息量进行度量, 具体如下:

$$H(X) = -\sum_{x \in X} p(x) \log p(x), \tag{18}$$

其中, $p(x)$ 是随机变量 X 的概率密度函数. 那么视频的信息量该如何度量呢? 本文参考文献 [244] 初步设想了视频信息量的度量方法.

首先, 为全世界所有的视频建立一个空间 Ω. 空间中任意两个视频的相似度代表两个视频的距离. 接下来, 以任意视频 $V_i \in \Omega$ [1)]为中心做一个单位超球, 超球中包含的视频个数就代表了该视频周围的密集程度 d_{V_i}. 显然拍摄生活中常见景象的视频周围应该比较密集, 而拍摄罕见景象的视频周围较为稀疏. 然而, 罕见景象的不确定性要远远高于常见景象, 进而所提供的信息量也较大. 比如: 绿色天空要远比蓝色天空罕见, 所引起的关注度也会更大. 基于此, 本文将视频空间中的密集程度进行归一化, 作为视频出现的概率, 如下:

$$p(V_i) = \frac{d_{V_i}}{\sum_{V \in \Omega} d_V}. \tag{19}$$

由此, 通过上式和式 (18) 可计算出整个视频空间的信息熵, 它实际上代表所有视频信息量的期望, 为避免符号混淆, 记为 $I(\Omega)$. 显然, 某个视频 V_i 所提供的信息量为

$$I(V_i) = -\log p(V_i). \tag{20}$$

如上所述, 本文依据经典信息论完成了对视频信息量的客观度量. 然而, 视频给人们带来信息的多少同时受观看者主观因素的影响. 例如, 一个登上过珠穆朗玛峰的人可能无法对普通登山视频产生新奇感, NBA 比赛的直播视频给篮球迷和其他人带来的感受也是不一样的. 因此, 本文倾向于认为同一视频对不同观看者带来的信息量并不相同, 有必要将观看者的主观因素加入到视频信息量的度量中, 定义如下:

$$I(V_i|\text{Viewer}) = -\log p(V_i|\text{Viewer}), \tag{21}$$

其中, 观看者 Viewer 的主观因素包括兴趣爱好、人生阅历、认知水平等.

(2) 视频空间的信容是多少? 问题 1 中粗略定义了视频空间, 这里对视频空间进一步精简. 为了简化问题, 将所有视频的空间尺度统一变换为 $W \times H$, 时间长度通过采样的方式统一为 T, 像素位深统一为 24 位, R, G, B 三个通道. 那么, 视频空间的理论视频数量

$$|\Omega| = 2^{W \times H \times T \times 24}. \tag{22}$$

1) 为了便于说明问题, 在不与前文产生矛盾的前提下, 本部分倾向于用标量符号表示所有变量.

每个视频的数据量是 $W \times H \times T \times 3$ 字节 (不考虑数据压缩). 那么整个视频空间的数据量为

$$D(\Omega) = W \times H \times T \times 3 \times 2^{W \times H \times T \times 24}. \tag{23}$$

进而, 整个视频空间的信容为

$$\psi(\Omega) = \frac{I(\Omega)}{D(\Omega)} = \frac{-\sum_{V \in \Omega} p(V) \log p(V)}{W \times H \times T \times 3 \times 2^{W \times H \times T \times 24}}. \tag{24}$$

类似地, 也可以对音频空间、图像空间、文本空间等的信容进行定义.

(3) 如何对视频空间进行萃取? 依据前文探讨的数据、信息与知识之间的关系, 人们从视频空间中所获取的知识为

$$\kappa(\Omega) = I(\Omega) \cdot \lambda. \tag{25}$$

显然, 任何人终其一生都无法遍历整个视频空间. 实际上, 视频空间如此庞大, 其中存在大量重复以及无意义的视频, 非常有必要对其进行萃取. 一方面, 人们日常生活中搜索到的视频存在大量重复剪辑的情况, 即同一镜头出现在多个视频中. 另一方面, 视频信息的传递需要考虑人的认知能力. 其中, (1) 随机像素值所生成的视频所有人都无法理解, 所以视频空间中某些理论存在的视频是绝对无意义的; (2) 不同人对同一视频的理解能力不同, 比如: 大学生对高等数学教学视频的理解能力要显著高于小学生, 所以某一视频对特定群体有可能是相对无意义的. 因此, 视频空间的萃取需要考虑个体的认知能力. 进一步地, 式 (25) 的描述严格来说并不准确, 因为随着人们不断地接收知识, 认知能力 λ 是逐渐提升的, 即

$$\lambda' = \lambda + \alpha I(V) \cdot \lambda, \tag{26}$$

其中, α 代表个体的学习能力. 总体来说, 视频空间萃取是视频萃取的延伸, 面临着和视频萃取类似的问题. 在考虑个体认知能力的前提下, 以在有限时间内获得更大的知识量为前提, 为不同人提供更加简洁的视频空间以及视频推荐系统, 是数据爆炸时代非常值得研究的课题.

(4) 信息萃取还有哪些用处? 随着科技的发展, 视频、图像、文本、音频等数据获取手段增多, 数据量呈现爆发式增长, 提高单位数据量的信息提供能力是各行各业针对各种数据类型的普遍需求. 通信行业中, 在保持原有信息量的同时去除冗余数据, 可以显著减少数据传输压力. 以视频会议为例, 参与人的背景信息往往是固定的, 只需传输一次, 可以大量缩减数据量, 同时目标信息得到保护. 教育行业中, 从书山题海中挑选出适合学生的学习资料, 可以提高学习效率, 同时将学生从繁杂的课业压力中解放出来. 医疗行业中, 信息萃取有助于链接各个医院的数据孤岛, 构建医疗大数据平台, 进行病例、临床、实验等方面的数据共享. 总体来说, 信息萃取可以用于各种数据类型, 推动各行各业的发展, 在大数据时代具有广泛的应用价值.

7 总结

本文从理论、方法和未来 3 个方面对视频萃取进行了细致的讨论分析. 首先, 从信息论的角度对视频萃取中数据、信息和知识的关系进行了理论解释, 创新性地统一了视频萃取各项任务的理论基础. 然后, 在对时空表征方法进行详尽的讨论与分析的基础上, 从视频摘要、视频浓缩、视频描述 3 个任务入手, 详细对比了从内容、目标和语义角度进行视频萃取的各种方法, 探讨了一系列科学问题以及所采取的解决方案, 全面综合地展示了视频萃取的历史沿革与发展现状, 发现了视频萃取发展中存在的一些关键问题. 最后, 本文分析了在视频数据爆炸时代视频萃取研究在数据建模、信息处理与个体

认知方面所面临的机遇与挑战, 并对未来在基础问题、数据处理与应用开发方面的发展趋势作出了展望. 希望本文能对视频萃取的研究提供理论、方法和应用方面的借鉴与启发, 吸引更多的科研工作者加入到相关研究中来, 在视频智能分析领域取得突破性进展.

参考文献

1 Li Q, Gkoumas D, Lioma C, et al. Quantum-inspired multimodal fusion for video sentiment analysis. Inf Fusion, 2021, 65: 58–71

2 Nie X S, Lin P G, Yang M Z, et al. Hierarchical feature fusion hashing for near-duplicate video retrieval. Sci Sin Inform, 2018, 48: 1697–1708 [聂秀山, 林培光, 杨明哲, 等. 基于层次特征融合哈希的近似重复视频检索方法. 中国科学: 信息科学, 2018, 48: 1697–1708]

3 Xiang S J, Yang J Q, Huang J W. Perceptual video Hashing robust against geometric distortions. Sci China Inf Sci, 2012, 55: 1520–1527 [项世军, 杨建权, 黄继武. 抗几何失真的视频 Hashing 算法研究. 中国科学: 信息科学, 2012, 42: 578–587]

4 Ghatak S, Rup S, Didwania H, et al. GAN based efficient foreground extraction and HGWOSA based optimization for video synopsis generation. Digital Signal Process, 2021, 111: 102988

5 Lokoc J, Soucek T, Vesely P, et al. A W2VV++ case study with automated and interactive text-to-video retrieval. In: Proceedings of the ACM International Conference on Multimedia, 2020. 2553–2561

6 Hu D, Nie F P, Li X L. Deep linear discriminant analysis hashing. Sci Sin Inform, 2021, 51: 279–293 [胡迪, 聂飞平, 李学龙. 基于深度线性判别分析的哈希技术. 中国科学: 信息科学, 2021, 51: 279–293]

7 Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text. In: Proceedings of the 18th National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, Edmonton, 2002. 786–791

8 Camastra F, Vinciarelli A. Machine Learning for Audio, Image and Video Analysis: Theory and Applications. London: Springer, 2015. 295–336

9 Li X, Zhang H, Zhang R, et al. Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning. IEEE Trans Image Process, 2020, 29: 2139–2149

10 Li X L, Gong H G. A survey on big data systems. Sci Sin Inform, 2015, 45: 1–44 [李学龙, 龚海刚. 大数据系统综述. 中国科学: 信息科学, 2015, 45: 1–44]

11 Li Z, Wang W, Li Z, et al. Towards visually explaining video understanding networks with perturbation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2021. 65–76

12 Aafaq N, Mian A, Liu W, et al. Video description: a survey of methods, datasets, and evaluation metrics. ACM Comput Surv, 2020, 52: 1–37

13 Li X, Song J, Gao L, et al. Beyond RNNs: positional self-attention with co-attention for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 8658–8665

14 Li X L, Zhao B, Lu X Q. Key frame extraction in the summary space. IEEE Trans Cybern, 2018, 48: 1923–1934

15 Schneider W J, Newman D A. Intelligence is multidimensional: theoretical review and implications of specific cognitive abilities. Human Resource Manage Rev, 2015, 25: 12–27

16 Hussain T, Muhammad K, Ding W, et al. A comprehensive survey of multi-view video summarization. Pattern Recogn, 2021, 109: 107567

17 Baskurt K B, Samet R. Video synopsis: a survey. Comput Vision Image Underst, 2019, 181: 26–38

18 Aafaq N, Akhtar N, Liu W, et al. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 12487–12496

19 Li X, Wang Z, Lu X. Video synopsis in complex situations. IEEE Trans Image Process, 2018, 27: 3798–3812

20 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. 886–893

21 Zhang K, Chao W, Sha F, et al. Video summarization with long short-term memory. In: Proceedings of the European Conference on Computer Vision, 2016. 766–782

22 Lowe D G. Distinctive image features from scale-invariant keypoints. Int J Comput Vision, 2004, 60: 91–110

23 Zhao B, Li X, Lu X. Hierarchical recurrent neural network for video summarization. In: Proceedings of the ACM Conference on Multimedia, 2017. 863–871

24 Venugopalan S, Rohrbach M, Donahue J, et al. Sequence to sequence — video to text. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4534–4542

25 Pan P, Xu Z, Yang Y, et al. Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1029–1038

26 Park H S, Jun C H. A simple and fast algorithm for K-medoids clustering. Expert Syst Appl, 2009, 36: 3336–3341

27 de Avila S E F, Lopes A P B, da Luz J A, et al. VSUMM: a mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn Lett, 2011, 32: 56–68

28 Papadopoulos D P, Kalogeiton V S, Chatzichristofis S A, et al. Automatic summarization and annotation of videos with lack of metadata information. Expert Syst Appl, 2013, 40: 5765–5778

29 Chu W, Song Y, Jaimes A. Video co-summarization: video summarization by visual co-occurrence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3584–3592

30 Elhamifar E, Sapiro G, Vidal R. See all by looking at a few: sparse modeling for finding representative objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012. 1600–1607

31 Mei S, Guan G, Wang Z, et al. L2,0 constrained sparse dictionary selection for video summarization. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 2014. 1–6

32 Ma M, Mei S, Wan S, et al. Video summarization via block sparse dictionary selection. Neurocomputing, 2020, 378: 197–209

33 Zhao B, Xing E P. Quasi real-time summarization for consumer videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 2513–2520

34 Gong B, Chao W, Grauman K, et al. Diverse sequential subset selection for supervised video summarization. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014. 2069–2077

35 Meng J, Wang H, Yuan J, et al. From keyframes to key objects: video summarization by representative object proposal selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1039–1048

36 Gygli M, Grabner H, van Gool L. Video summarization by learning submodular mixtures of objectives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3090–3098

37 Pan J, Yang H, Faloutsos C. MMSS: multi-modal story-oriented video summarization. In: Proceedings of the IEEE International Conference on Data Mining, 2004. 491–494

38 Li X, Zhao B, Lu X. A general framework for edited video and raw video summarization. IEEE Trans Image Process, 2017, 26: 3652–3664

39 Zhao B, Li X, Lu X. Property-constrained dual learning for video summarization. IEEE Trans Neural Netw Learn Syst, 2020, 31: 3989–4000

40 Lee Y J, Ghosh J, Grauman K. Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012. 1346–1353

41 Pritch Y, Rav-Acha A, Peleg S. Nonchronological video synopsis and indexing. IEEE Trans Pattern Anal Mach Intell, 2008, 30: 1971–1984

42 Jin J, Liu F, Gan Z, et al. Online video synopsis method through simple tube projection strategy. In: Proceedings of the 8th International Conference on Wireless Communications & Signal Processing, 2016. 1–5

43 Hoshen Y, Peleg S. Live video synopsis for multiple cameras. In: Proceedings of 2015 IEEE International Conference on Image Processing, 2015. 212–216

44 Taj M, Maggio E, Cavallaro A. Multi-feature graph-based object tracking. In: Proceedings of Multimodal Technologies for Perception of Humans, First International Evaluation Workshop on Classification of Events, Activities and Relationships, 2006. 190–199

45 He Y, Qu Z, Gao C, et al. Fast online video synopsis based on potential collision graph. IEEE Signal Process Lett, 2017, 24: 22–26

46 Sun L, Xing J, Ai H, et al. A tracking based fast online complete video synopsis approach. In: Proceedings of the 21st International Conference on Pattern Recognition, 2012. 1956–1959

47 Hsia C H, Chiang J S, Hsieh C F. Low-complexity range tree for video synopsis system. Multimed Tools Appl, 2016,

75: 9885–9902
48 Zhu X, Liu J, Wang J, et al. Key observation selection-based effective video synopsis for camera network. Machine Vision Appl, 2014, 25: 145–157
49 Zhu X, Loy C C, Gong S. Learning from multiple sources for video summarisation. Int J Comput Vis, 2016, 117: 247–268
50 Lin L, Lin W, Xiao W, et al. An optimized video synopsis algorithm and its distributed processing model. Soft Comput, 2017, 21: 935–947
51 Yan C, Tu Y, Wang X, et al. STAT: spatial-temporal attention mechanism for video captioning. IEEE Trans Multimedia, 2020, 22: 229–241
52 Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 1627–1645
53 Moore D J, Essa I A. Recognizing multitasked activities from video using stochastic context-free grammar. In: Proceedings of the 18th National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, 2002. 770–776
54 Zhu S C, Mumford D. A stochastic grammar of images. FNT Comput Graph Vision, 2006, 2: 259–362
55 Donahue J, Hendricks L A, Rohrbach M, et al. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 677–691
56 Zhao B, Li X, Lu X. Video captioning with tube features. In: Proceedings of International Joint Conference on Artificial Intelligence, 2018. 1177–1183
57 Sah S, Nguyen T, Ptucha R. Understanding temporal structure for video captioning. Pattern Anal Applic, 2020, 23: 147–159
58 Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4507–4515
59 Li X, Zhao B, Lu X. MAM-RNN: multi-level attention model based RNN for video captioning. In: Proceedings of the International Joint Conference on Artificial Intelligence, 2017. 2208–2214
60 Li L, Gong B. End-to-end video captioning with multitask reinforcement learning. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2019. 339–348
61 Dai B, Fidler S, Urtasun R, et al. Towards diverse and natural image descriptions via a conditional GAN. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 2989–2998
62 Rahman T, Xu B, Sigal L. Watch, listen and tell: multi-modal weakly supervised dense event captioning. In: Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, 2019. 8907–8916
63 Lu C, Shi J, Wang W, et al. Fast abnormal event detection. Int J Comput Vis, 2019, 127: 993–1011
64 Basavarajaiah M, Sharma P. Survey of compressed domain video summarization techniques. ACM Comput Surv, 2020, 52: 1–29
65 Zhang Y, Liang X, Zhang D, et al. Unsupervised object-level video summarization with online motion auto-encoder. Pattern Recognition Lett, 2020, 130: 376–385
66 Wang W, Shen J, Lu X, et al. Paying attention to video object pattern understanding. IEEE Trans Pattern Anal Mach Intell, 2020. doi: 10.1109/TPAMI.2020.2966453
67 Wan S, Goudos S. Faster R-CNN for multi-class fruit detection using a robotic vision system. Comput Networks, 2020, 168: 107036
68 Li X L, Shi J H, Dong Y S, et al. A survey on scene image classification. Sci Sin Inform, 2015, 45: 827–848 [李学龙, 史建华, 董永生, 等. 场景图像分类技术综述. 中国科学: 信息科学, 2015, 45: 827–848]
69 Fooladgar F, Kasaei S. A survey on indoor RGB-D semantic segmentation: from hand-crafted features to deep convolutional neural networks. Multimed Tools Appl, 2020, 79: 4499–4524
70 Han J, Ma K-K. Fuzzy color histogram and its use in color image retrieval. IEEE Trans Image Process, 2002, 11: 944–952
71 Bay H, Tuytelaars T, van Gool L. SURF: speeded up robust features. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2006. 404–417
72 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012. 1106–1114

73 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations, 2015

74 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3431–3440

75 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 770–778

76 Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1–9

77 Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4700–4708

78 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255

79 Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale. 2018. ArXiv:1811.00982

80 Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proc IEEE, 1998, 86: 2278–2324

81 Krizhevsky A, Hinton G, et al. Learning multiple layers of features from tiny images. 2009. 32–35

82 Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database. In: Proceedings of Advances in Neural Information Processing Systems, 2014. 487–495

83 Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human actions classes from videos in the wild. 2012. ArXiv:1212.0402

84 Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition. In: Proceedings of 2011 International Conference on Computer Vision, 2011. 2556–2563

85 Monfort M, Andonian A, Zhou B, et al. Moments in time dataset: one million videos for event understanding. 2018. ArXiv:1801.03150

86 Heilbron C, Escorcia V, Ghanem B, et al. ActivityNet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 961–970

87 Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. 2017. ArXiv:1705.06950

88 Tan X Y, Triggs B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans Image Process, 2010, 19: 1635–1650

89 Smith J R, Chang S. Automated binary texture feature sets for image retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. 2239–2242

90 Song Y, Vallmitjana J, Stent A, et al. TVSum: summarizing web videos using titles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 5179–5187

91 Hafner J, Sawhney H S, Equitz W, et al. Efficient color histogram indexing for quadratic form distance functions. IEEE Trans Pattern Anal Machine Intell, 1995, 17: 729–736

92 Li W X, Lou X P, Dong M L, et al. Golf video tracking based on recognition with HOG and spatial-temporal vector. Int J Adv Robotic Syst, 2017, 14: 172988141770454

93 Kumar K P S, Bhavani R. Human activity recognition in egocentric video using HOG, GiST and color features. Multimed Tools Appl, 2020, 79: 3543–3559

94 Lee T, Hwangbo M, Alan T, et al. Low-complexity HOG for efficient video saliency. In: Proceedings of 2015 IEEE International Conference on Image Processing (ICIP), 2015. 3749–3752

95 Zhu Y, Huang X, Huang Q, et al. Large-scale video copy retrieval with temporal-concentration SIFT. Neurocomputing, 2016, 187: 83–91

96 Battiato S, Gallo G, Puglisi G, et al. SIFT features tracking for video stabilization. In: Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP 2007), 2007. 825–830

97 Li X L, Chen M L, Wang Q. Multiview-based group behavior analysis in optical image sequence. Sci Sin Inform, 2018, 48: 1227–1241 [李学龙, 陈穆林, 王琦. 光学影像序列中基于多视角聚类的群组行为分析. 中国科学: 信息科学, 2018, 48: 1227–1241]

98 Mhaskar H N, Poggio T. Deep vs. shallow networks: an approximation theory perspective. Anal Appl, 2016, 14:

829–848
99 Ren Z, Xu W. An improved path integration method for nonlinear systems under Poisson white noise excitation. Appl Math Comput, 2020, 373: 125036
100 Zhou K, Qiao Y, Xiang T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2018. 7582–7589
101 Yuan Y, Li H, Wang Q. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. IEEE Access, 2019, 7: 64676–64685
102 Yao T, Mei T, Rui Y. Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 982–990
103 Michelucci U. Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection. New York: Apress, 2019
104 Gibson J, Marques O. Optical Flow and Trajectory Estimation Methods. Berlin: Springer, 2019. 9–21
105 Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks. In: Proceedings of IEEE International Conference on Computer Vision, 2015. 2758–2766
106 Ilg E, Mayer N, Saikia T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1647–1655
107 Ranjan A, Black M J. Optical flow estimation using a spatial pyramid network. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2720–2729
108 Sun D, Yang X, Liu M, et al. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 8934–8943
109 Revaud J, Weinzaepfel P, Harchaoui Z, et al. EpicFlow: edge-preserving interpolation of correspondences for optical flow. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1164–1172
110 Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. In: Proceedings of the 27th International Conference on Machine Learning, 2010. 495–502
111 Tran D, Bourdev L D, Fergus R, et al. C3D: generic features for video analysis. 2014. ArXiv:1412.0767
112 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 4724–4733
113 Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of IEEE International Conference on Computer Vision Workshops, 2017. 3154–3160
114 Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 5534–5542
115 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780
116 Cho K, van Merrienboer B, Bahdanau D, et al. On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014. 103–111
117 Zhao B, Li X, Lu X. HSA-RNN: hierarchical structure-adaptive RNN for video summarization. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7405–7414
118 Arabshahi F, Lu Z, Singh S, et al. Memory augmented recursive neural networks. 2019. ArXiv:1911.03329
119 Ren Z, Xu W, Zhang S. Reliability analysis of nonlinear vibro-impact systems with both randomly fluctuating restoring and damping terms. Commun Nonlin Sci Numer Simul, 2020, 82: 105087
120 Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2982–2991
121 Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw, 1994, 5: 157–166
122 Hu Z, Nie F, Wang R, et al. Multi-view spectral clustering via integrating nonnegative embedding and spectral embedding. Inf Fusion, 2020, 55: 251–259
123 Nie F, Hu Z, Li X. Matrix completion based on non-convex low-rank approximation. IEEE Trans Image Process, 2019, 28: 2378–2388
124 Chinrungrueng C, Sequin C H. Optimal adaptive k-means algorithm with dynamic adjustment of learning rate. IEEE Trans Neural Netw, 1995, 6: 157–169

125 Frey B J, Dueck D. Clustering by passing messages between data points. Science, 2007, 315: 972–976

126 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. Science, 2014, 344: 1492–1496

127 Ren J, Jiang J, Feng Y. Activity-driven content adaptation for effective video summarization. J Visual Commun Image Represent, 2010, 21: 930–938

128 Khosla A, Hamid R, Lin C, et al. Large-scale video summarization using web-image priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. 2698–2705

129 Cong Y, Yuan J, Luo J. Towards scalable summarization of consumer videos via sparse dictionary selection. IEEE Trans Multimedia, 2012, 14: 66–75

130 Gong Y, Liu X. Video summarization using singular value decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000. 174–180

131 Wang S, Cong Y, Cao J, et al. Scalable gastroscopic video summarization via similar-inhibition dictionary selection. Artificial Intell Med, 2016, 66: 1–13

132 Etezadifar P, Farsi H. Scalable video summarization via sparse dictionary learning and selection simultaneously. Multimed Tools Appl, 2017, 76: 7947–7971

133 Marvaniya S, Damoder M, Gopalakrishnan V, et al. Real-time video summarization on mobile. In: Proceedings of IEEE International Conference on Image Processing, 2016. 176–180

134 Wang J, Wang Y, Zhang Z. Visual saliency based aerial video summarization by online scene classification. In: Proceedings of International Conference on Image and Graphics, 2011. 777–782

135 Yao T, Mei T, Rui Y. Highlight detection with pairwise deep ranking for first-person video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 982–990

136 Atencio P, German S T, Branch J W, et al. Video summarisation by deep visual and categorical diversity. IET Comput Vision, 2019, 13: 569–577

137 Kaushal V, Iyer R, Doctor K, et al. Demystifying multi-faceted video summarization: tradeoff between diversity, representation, coverage and importance. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2019. 452–461

138 Ejaz N, Tariq T B, Baik S W. Adaptive key frame extraction for video summarization using an aggregation mechanism. J Visual Commun Image Represent, 2012, 23: 1031–1040

139 Macchi O. The coincidence approach to stochastic point processes. Adv Appl Probab, 1975, 7: 83–122

140 Liu D, Hua G, Chen T. A hierarchical visual model for video object summarization. IEEE Trans Pattern Anal Mach Intell, 2010, 32: 2178–2190

141 Li X, Chen M, Nie F, et al. A multiview-based parameter free framework for group detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2017. 4147–4153

142 Lee Y J, Grauman K. Predicting important objects for egocentric video summarization. Int J Comput Vis, 2015, 114: 38–55

143 Yang L, Cheng H, Su J, et al. Pixel-to-model distance for robust background reconstruction. IEEE Trans Circ Syst Video Technol, 2016, 26: 903–916

144 Liu Q, Fang L, Yu G, et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nature Commun, 2019, 10: 1–11

145 Zhao B, Li X, Lu X. TTH-RNN: tensor-train hierarchical recurrent neural network for video summarization. IEEE Trans Indust Electron, 2021, 68: 3629–3637

146 Alemany S, Beltran J, Perez A, et al. Predicting hurricane trajectories using a recurrent neural network. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 468–475

147 Yang H, Wang B, Lin S, et al. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 4633–4641

148 Zhang K, Chao W, Sha F, et al. Summary transfer: exemplar-based subset selection for video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 1059–1067

149 Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 2982–2991

150 Apostolidis E E, Adamantidou E, Metsai A I, et al. Unsupervised video summarization via attention-driven adversarial learning. In: Proceedings of the International Conference on Multimedia Modeling, 2020. 492–504

151 Ji Z, Xiong K, Pang Y, et al. Video summarization with attention-based encoder-decoder networks. IEEE Trans Circ Syst Video Technol, 2020, 30: 1709–1717
152 Rochan M, Ye L, Wang Y. Video summarization using fully convolutional sequence networks. In: Proceedings of the European Conference on Computer Vision, 2018. 347–363
153 Rochan M, Wang Y. Video summarization by learning from unpaired data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7902–7911
154 Potapov D, Douze M, Harchaoui Z, et al. Category-specific video summarization. In: Proceedings of the European Conference on Computer Vision, 2014. 540–555
155 Gygli M, Grabner H, Riemenschneider H, et al. Creating summaries from user videos. In: Proceedings of the European Conference on Computer Vision, 2014. 505–520
156 Zeng K, Chen T, Niebles J C, et al. Title generation for user generated videos. In: Proceedings of the European Conference on Computer Vision, 2016. 609–625
157 Meng J, Wang S, Wang H, et al. Video summarization via multi-view representative selection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1189–1198
158 Anirudh R, Masroor A, Turaga P K. Diversity promoting online sampling for streaming video summarization. In: Proceedings of the IEEE International Conference on Image Processing, 2016. 3329–3333
159 Otani M, Nakashima Y, Rahtu E, et al. Rethinking the evaluation of video summaries. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 7596–7604
160 Rav-Acha A, Pritch Y, Peleg S. Making a long video short: dynamic video synopsis. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. 1–7
161 Ra M, Kim W Y. Parallelized tube rearrangement algorithm for online video synopsis. IEEE Signal Process Lett, 2018, 25: 1186–1190
162 Ghatak S, Rup S, Majhi B, et al. An improved surveillance video synopsis framework: a HSATLBO optimization approach. Multimed Tools Appl, 2020, 79: 4429–4461
163 Nie Y, Xiao C, Sun H, et al. Compact video synopsis via global spatiotemporal optimization. IEEE Trans Visual Comput Graph, 2013, 19: 1664–1676
164 Wang S, Xu W, Chao W, et al. A framework for surveillance video fast browsing based on object flags. In: Proceedings of Pacific-rim Conference on Multimedia, 2013. 411–421
165 Feng S, Lei Z, Yi D, et al. Online content-aware video condensation. In: Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012. 2082–2087
166 Huang C R, Chung P C J, Yang D K, et al. Maximum a posteriori probability estimation for online surveillance video synopsis. IEEE Trans Circ Syst Video Technol, 2014, 24: 1417–1429
167 Lu M, Wang Y, Pan G. Generating fluent tubes in video synopsis. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2013. 2292–2296
168 Chou C, Lin C, Chiang T, et al. Coherent event-based surveillance video synopsis using trajectory clustering. In: Proceedings of 2015 IEEE International Conference on Multimedia & Expo Workshops, 2015. 1–6
169 Zhu X, Loy C C, Gong S. Video synopsis by heterogeneous multi-source correlation. In: Proceedings of IEEE International Conference on Computer Vision, 2013. 81–88
170 Nie F, Wang Z, Wang R, et al. Towards robust discriminative projections learning via non-greedy $l_{2,1}$-norm minmax. IEEE Trans Pattern Anal Mach Intell, 2020. doi: 10.1109/TPAMI.2019.2961877
171 Mahapatra A, Sa P K, Majhi B, et al. MVS: a multi-view video synopsis framework. Signal Process Image Commun, 2016, 42: 31–44
172 Lin W, Zhang Y, Lu J, et al. Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. Neurocomputing, 2015, 155: 84–98
173 Pritch Y, Rav-Acha A, Gutman A, et al. Webcam synopsis: peeking around the world. In: Proceedings of IEEE International Conference on Computer Vision, 2007. 1–8
174 Correa C D, Ma K. Dynamic video narratives. ACM Trans Graph, 2010, 29: 1–9
175 Xu M, Li S Z, Li B, et al. A set theoretical method for video synopsis. In: Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, 2008. 366–370
176 Li X, Wang Z, Lu X. Surveillance video synopsis via scaling down objects. IEEE Trans Image Process, 2016, 25:

740–755
177 Nie Y, Li Z, Zhang Z, et al. Collision-free video synopsis incorporating object speed and size changes. IEEE Trans Image Process, 2020, 29: 1465–1478
178 He Y, Gao C, Sang N, et al. Graph coloring based surveillance video synopsis. Neurocomputing, 2017, 225: 64–79
179 Yildiz A, Ozgur A, Akgul Y S. Fast non-linear video synopsis. In: Proceedings of International Symposium on Computer and Information Sciences, 2008. 1–6
180 Vural U, Akgul Y S. Eye-gaze based real-time surveillance video synopsis. Pattern Recognition Lett, 2009, 30: 1151–1159
181 Kirkpatrick S, Gelatt C D, Vecchi M P. Optimization by simulated annealing. Science, 1983, 220: 671–680
182 Ghatak S, Rup S, Majhi B, et al. HSAJAYA: an improved optimization scheme for consumer surveillance video synopsis generation. IEEE Trans Consumer Electron, 2020, 66: 144–152
183 Rao R V. Jaya: a simple and new optimization algorithm for solving constrained and unconstrained optimization problems. Int J Indust Eng Comput, 2016, 7: 19–34
184 Ruan T, Wei S, Li J, et al. Rearranging online tubes for streaming video synopsis: a dynamic graph coloring approach. IEEE Trans Image Process, 2019, 28: 3873–3884
185 Liao W, Tu Z, Wang S, et al. Compressed-domain video synopsis via 3D graph cut and blank frame deletion. In: Proceedings of the on Thematic Workshops of ACM Multimedia, 2017. 253–261
186 Zhong R, Hu R M, Wang Z Y, et al. Fast synopsis for moving objects using compressed video. IEEE Signal Process Lett, 2014, 21: 834–838
187 Zhang Z, Nie Y, Sun H, et al. Multi-view video synopsis via simultaneous object-shifting and view-switching optimization. IEEE Trans Image Process, 2020, 29: 971–985
188 Zhao Z Q, Zheng P, Xu S T, et al. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst, 2019, 30: 3212–3232
189 Chin T, Ding R, Marculescu D. Adascale: towards real-time video object detection using adaptive scaling. 2019. ArXiv:1902.02910
190 Alwando E H P, Chen Y T, Fang W H. CNN-based multiple path search for action tube detection in videos. IEEE Trans Circ Syst Video Technol, 2020, 30: 104–116
191 Yuan Y, Wang D, Wang Q. Memory-augmented temporal dynamic learning for action recognition. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, 2019. 9167–9175
192 Li X, Chen M, Nie F, et al. Locality adaptive discriminant analysis. In: Proceedings of International Joint Conference on Artificial Intelligence, 2017. 2201–2207
193 Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 2013. 2712–2719
194 Thomason J, Venugopalan S, Guadarrama S, et al. Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of the International Conference on Computational Linguistics, 2014. 1218–1227
195 Krishnamoorthy N, Malkarnenkar G, Mooney R J, et al. Generating natural-language video descriptions using text-mined knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2013
196 Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vis, 2002, 50: 171–184
197 Gong S, Xiang T. Recognition of group activities using dynamic probabilistic networks. In: Proceedings of the 9th IEEE International Conference on Computer Vision, 2003. 742–749
198 Bobick A F, Wilson A D. A state-based approach to the representation and recognition of gesture. IEEE Trans Pattern Anal Machine Intell, 1997, 19: 1325–1337
199 Hanckmann P, Schutte K, Burghouts G J. Automated textual descriptions for a wide range of video events with 48 human actions. In: Computer Vision — ECCV 2012. Berlin: Springer, 2012. 372–380
200 Pan Y, Yao T, Li H, et al. Video captioning with transferred semantic attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 984–992
201 Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3185–3194
202 Cherian A, Wang J, Hori C, et al. Spatio-temporal ranked-attention networks for video captioning. 2020. ArXiv:2001.06127
203 Nabati M, Behrad A. Video captioning using boosted and parallel long short-term memory networks. Comput Vision Image Underst, 2020, 190: 102840
204 Guo Y, Zhang J, Gao L. Exploiting long-term temporal dynamics for video captioning. World Wide Web, 2019, 22: 735–749
205 Wang H, Gao C, Han Y. Sequence in sequence for video captioning. Pattern Recogn Lett, 2020, 130: 327–334
206 Venugopalan S, Xu H, Donahue J, et al. Translating videos to natural language using deep recurrent neural networks. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015. 1494–1504
207 Long X, Gan C, de Melo G. Video captioning with multi-faceted attention. TACL, 2018, 6: 173–184
208 Yu Y, Ko H, Choi J, et al. Video captioning and retrieval models with semantic attention. 2016. ArXiv:1610.02947
209 Venugopalan S, Hendricks L A, Mooney R J, et al. Improving lSTM-based video description with linguistic knowledge mined from text. In: Proceedings of Conference on Empirical Methods in Natural Language Processing, 2016. 1961–1966
210 Zhao B, Li X, Lu X. CAM-RNN: co-attention model based RNN for video captioning. IEEE Trans Image Process, 2019, 28: 5552–5565
211 Hori C, Hori T, Lee T, et al. Attention-based multimodal fusion for video description. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 4203–4212
212 Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3185–3194
213 Ren L, Qi G J, Hua K. Improving diversity of image captioning through variational autoencoders and adversarial learning. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2019. 263–272
214 Park J S, Rohrbach M, Darrell T, et al. Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6598–6608
215 Ren Z, Wang X, Zhang N, et al. Deep reinforcement learning-based image captioning with embedding reward. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1151–1159
216 Wang X, Chen W, Wu J, et al. Video captioning via hierarchical reinforcement learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 4213–4222
217 Zhang W, Wang B, Ma L, et al. Reconstruct and represent video contents for captioning via reinforcement learning. 2019. ArXiv:1906.01452
218 Song J, Guo Y, Gao L, et al. From deterministic to generative: multimodal stochastic RNNs for video captioning. IEEE Trans Neural Netw Learn Syst, 2019, 30: 3047–3058
219 Wu A, Han Y. Hierarchical memory decoding for video captioning. 2020. ArXiv:2002.11886
220 Chen D L, Dolan W B. Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011. 190–200
221 Regneri M, Rohrbach M, Wetzel D, et al. Grounding action descriptions in videos. Trans Assoc Comput Linguist, 2013, 1: 25–36
222 Torabi A, Pal C J, Larochelle H, et al. Using descriptive video services to create a large data source for video annotation research. 2015. ArXiv:1503.01070
223 Rohrbach A, Rohrbach M, Tandon N, et al. A dataset for movie description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 3202–3212
224 Xu J, Mei T, Yao T, et al. MSR-VTT: a large video description dataset for bridging video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 5288–5296
225 Sigurdsson G A, Varol G, Wang X, et al. Hollywood in homes: crowdsourcing data collection for activity understanding. In: Proceedings of the European Conference on Computer Vision, 2016. 510–526
226 Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002. 311–318
227 Lin C, Och F J. Automatic evaluation of machine translation quality using longest common subsequence and skip-

bigram statistics. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2004. 605–612

228 Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 4566–4575

229 Denkowski M J, Lavie A. METEOR universal: language specific translation evaluation for any target language. In: Proceedings of the Workshop on Statistical Machine Translation, 2014. 376–380

230 Xu J, Yao T, Zhang Y, et al. Learning multimodal attention LSTM networks for video captioning. In: Proceedings of the ACM International Conference on Multimedia, 2017. 537–545

231 Yu H, Wang J, Huang Z, et al. Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016. 4584–4593

232 Liu S, Ren Z, Yuan J. SibNet: sibling convolutional encoder for video captioning. IEEE Trans Pattern Anal Machine Intell, 2020. doi: 10.1109/TPAMI.2019.2940007

233 Ballas N, Yao L, Pal C, et al. Delving deeper into convolutional networks for learning video representations. In: Proceedings of International Conference on Learning Representations, 2016. 1–11

234 Chen Y, Wang S, Zhang W, et al. Less is more: picking informative frames for video captioning. In: Proceedings of European Conference on Computer Vision, 2018. 367–384

235 Gan Z, Gan C, He X, et al. Semantic compositional networks for visual captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1141–1150

236 Wang B, Ma L, Zhang W, et al. Reconstruction network for video captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7622–7631

237 Wang J, Wang W, Huang Y, et al. M3: multimodal memory modelling for video captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 7512–7520

238 Krishna R, Hata K, Ren F, et al. Dense-captioning events in videos. In: Proceedings of IEEE International Conference on Computer Vision, 2017. 706–715

239 Zhang J, Peng Y. Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 8327–8336

240 Zhang X, Gao K, Zhang Y, et al. Task-driven dynamic fusion: reducing ambiguity in video description. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3713–3721

241 Shetty R, Laaksonen J. Frame-and segment-level features and candidate pool evaluation for video caption generation. In: Proceedings of ACM Conference on Multimedia Conference, 2016. 1073–1076

242 Hori C, Hori T, Lee T, et al. Attention-based multimodal fusion for video description. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. 4203–4212

243 Mun J, Yang L, Ren Z, et al. Streamlined dense video captioning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2019. 6588–6597

244 Li X L. Features, indexing, and interaction in content-based image retrieval. Dissertation for Ph.D. Degree. Hefei: University of Science and Technology of China, 2002 [李学龙. 基于内容的图像检索中特征、索引及交互问题研究. 博士学位论文. 合肥: 中国科学技术大学, 2002]

Video distillation

Xuelong LI* & Bin ZHAO

School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072*, China*
* Corresponding author. E-mail: li@nwpu.edu.cn

Abstract Video has become one of the most important data forms. Video distillation explores more compact data forms and information modalities by analyzing the spatial-temporal and semantic features of video data, which is an important task in computer vision and a key technique in artificial intelligence. With the rapid development of video capturing devices and the increasing human requirements, video analysis tasks are facing numbers of opportunities and challenges. In recent years, large amounts of video distillation approaches are proposed. This paper creatively unifies the theoretical basis of video distillation by analyzing the relationship among data, information and knowledge from the perspective of information theory, and argues that the principle of video distillation is to improve the information capacity of video data. Then, we overview existing approaches in the aspects of video data representation, key content summarization, moving object synopsis and text description generation, etc., and relate the development of video summarization, synopsis and captioning, which are typical tasks in video distillation. More importantly, this paper discusses the advantages and drawbacks of existing approaches, and then points out several key scientific problems that have not yet been addressed, and simultaneously analyzes the potential future development in video distillation.

Keywords video distillation, visual representation, video summarization, video synopsis, video captioning, computer vision, artificial intelligence

Xuelong LI was born in 1976. He received his Ph.D. degree in electronic engineering and information science from University of Science and Technology of China, Hefei, China, in 2002. He is currently a full professor at the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China.

Bin ZHAO was born in 1993. He received his Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently a post-doctor in Academy of Advanced Interdisciplinary Research, Xidian University, Xi'an, China. He focuses on introducing physical and cognitive models to artificial intelligence.