引用格式：董柏岩，王树祺，金鑫.视频精彩集锦生成技术综述[J].中国传媒大学学报（自然科学版），2022，29(02)：63-69.
文章编号：1673-4793(2022)02-0063-07

视频精彩集锦生成技术综述

董柏岩[1]，王树祺[2*]，金鑫[1]

（1.北京电子科技学院，北京 100070；2.国家开发投资集团有限公司，北京 100034）

摘要：由于便携拍摄设备的普及，以及社交媒体平台的推广，互联网上的视频数量呈爆炸式增长。视频精彩集锦生成技术可以分析给定的视频内容，自动剪辑出视频中的精彩片段，减轻人工处理视频的资源压力和成本。本文整理了视频精彩集锦生成技术的相关研究工作，介绍了有监督学习方法和无监督学习方法，分析了这些技术的优缺点，最后介绍了其在现实生活中的应用价值。
关键词：视频精彩集锦生成；有监督学习；无监督学习
中图分类号：TP391　　**文献标识码**：A

Overview of video highlights generation technology

DONG Boyan[1]，WANG Shuqi[2*]，JIN Xin[1]

(1. Beijing Electronic Science and Technology Institute, Beijing 100070, China;
2. State Development & Investment Corp LTD, Beijing 100034, China)

Abstract: Due to the popularity of portable shooting devices and the promotion of social media platforms, the number of videos on the Internet has increased explosively. Video highlights generation technology can analyze the given video content, automatically clip the highlights in the video, and reduce the resource pressure and cost of manual video processing. This paper sorts out the relevant research work of video highlights generation technology, introduces supervised learning methods and unsupervised learning methods, analyzes the advantages and disadvantages of these technologies, and finally introduces their application value in real life.
Key words: video highlights generation; supervised learning; unsupervised learning

1　引言

通讯技术和互联网技术的快速发展改变了人们生活的方方面面，人们已习惯于拍摄各种内容丰富的视频记录和分享生活，便携式拍摄设备和社交媒体平台的推广则使互联网上的视频数量呈现了爆炸式的增长。然而，处理这些海量的视频需要花费巨量的人力物力资源，为了缓解因视频数量增长而不断加大的数据处理压力，学术界开始研究视频精彩集锦生成技术。视频精彩集锦生成技术的目的是从一段完整的视频中自动选择最具有吸引力、最让人们感兴趣的一部分。这种技术一方面可以节省人们观看视频的时间，提高观看感受；另一方面使视频平台可以通过推荐精彩镜头来提高视频的吸引力，引导他人观看完整视频。因此，视频精彩集锦生成技术在多个领域均有着重要的应用价值。

作者简介(*为通讯作者)：董柏岩(1998-)，男，硕士研究生，主要研究方向：人工智能；王树祺(1986-)，男，工程师，硕士，主要研究方向：档案信息化、模式识别与智能系统、网络安全。Email:wangshuqi@sdic.com.cn；金鑫(1983-)，男，副教授，博士，主要研究方向：计算美学、计算机视觉、人工智能与安全。Email：jinxin@besti.edu.cn

最早的关于视频精彩集锦生成技术研究集中在体育视频的剪辑[1-4],近年来,研究的主题更加丰富,研究者们开始研究互联网视频[5]和第一人称视频[6],提出了许多新颖的视频精彩集锦生成方法。虽然研究视频的主题有所拓展,但这些方法大都只能应用于特定领域,即可以使用这些方法的视频大都有着相同的主题,如足球、滑雪等。这说明对不同主题的视频,"精彩"的定义也不相同。

现有的视频精彩集锦生成方法主要遵循两种策略。第一种策略将视频精彩集锦生成视为一项有监督学习任务[5-7]。人们对没有经过剪辑的视频进行人工标注,标记视频的精彩片段作为训练数据进行训练,使视频中的精彩部分获得更高的分数。虽然按这种方式设计的视频精彩集锦生成方法具有较好的性能,可以良好的识别视频的精彩片段,但这种方法工作量大,且难以拓展。第二种策略将视频精彩集锦生成视为弱监督或无监督的识别任务[8-10]。给定一特定领域的视频,视频精彩集锦生成方法会发现在训练样本中经常出现的内容,并学会在同一领域的新视频中检测这些片段作为精彩集锦。这种方法在监督方面具有可拓展性,能够利用视频时长等信息对精彩片段进行检测,缺点是辨别力不强,即样本之间的重复并不代表片段的精彩程度高。两种策略各有优劣,研究者们所提出的视频精彩集锦生成方法大都属于这两种策略。

本文后面章节将分别介绍基于有监督学习和无监督/弱监督学习的视频精彩集锦生成方法,并分析这些方法的优劣之处,最后介绍视频精彩集锦生成技术的应用价值与意义。

2 有监督学习方法

有监督学习指通过已有的训练样本去训练得到一个最优模型,再利用这个模型将所有的输入映射为相应的输出。对于视频精彩集锦生成任务而言,训练样本即视频和人工标注的视频精彩片段,标注好的精彩片段比视频的其他片段有着更高的分数,在排序中排名靠前。有监督的视频精彩集锦生成方法是数据驱动的,因此它们的性能高度依赖于人类标记的训练数据。一般来说,基于有监督学习的视频精彩集锦生成技术有着较好的性能,缺点是由于性能高度依赖于训练数据,导致方法的拓展性、通用性不强,且生成训练数据需要大量的时间和精力。

2.1 传统方法

早期的有监督学习方法利用了视频的视听特征[1,12]和视觉语义[11]。Rui等[1]研究了棒球比赛的精彩集锦生成问题,并提出了一种仅使用音频特征进行精彩片段检测的方法。使用的音频特征包括了能量相关特征、音素级特征、信息复杂性特征和韵律特征等。这些特征被设计用来解决不同的问题,如使用音素级特征中的梅尔频率倒谱系数来分辨人类语音。由于仅使用了音频特征,这种方法所需的计算力较少,即使在本地机顶盒上也可以进行集锦的生成。Rui等假设棒球比赛的精彩部分在投球和击球之后且播音员激动的解说高度相关,因此提出的精彩集锦生成算法先检测人类兴奋时的语音和棒球击球声,然后智能地融合它们以生成最终的精彩集锦。由于棒球比赛过程中包含了多种噪音,Rui等还开发了噪声环境下鲁棒的语音端点检测技术,并将支持向量机应用于语音分类。算法流程图见图1。

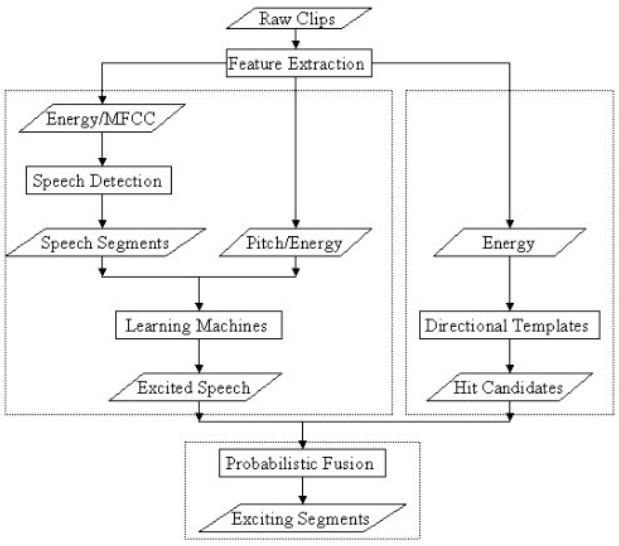

图1 Rui等提出的棒球比赛集锦生成算法流程图

2.2 深度学习方法

而最新的方法则基于深度学习构建视频精彩集锦生成模型[6,13],这些模型训练了多层神经网络来预测输入视频片段的精彩程度。模型的输入是视频片段的紧凑表示(如视频帧经过卷积得到的视觉特征),输出是一个标量值,以分数的形式表示,代表着输入视频的精彩程度。训练时,通过排序损失函数对神经网络进行训练,使视频精彩部分的得分高于其他部分的得分。在测试阶段,经过训练的模型可以预测任何

输入视频的精彩程度。

2.2.1　基于双流神经网络的方法

Yao 等[6]研究了第一人称视频的精彩集锦生成问题，提出了一种成对深度排名模型，该模型采用深度学习技术来学习视频精彩片段和非精彩片段之间的关系。Yao 等的精彩集锦生成方法流程如下，首先将输入视频分割为一组片段，每个视频片段被分解为空间和时间流，空间流以帧的形式出现，而时间流以视频片段的格式表示，一种用于精彩镜头预测的双流深度卷积神经网络结构被设计并用于空间流和时间流。这两个分量的输出通过后期融合进行组合，作为每个视频片段的精彩程度分数。分数高的片段就是原视频中的精彩部分，根据精彩分数便可以生成视频的精彩集锦。方法框架如图 2 所示。Yao 等还构建了一个新的数据集，内容包括了 15个体育相关主题，每一主题有大约 40个视频，视频长度在 2 到 15分钟之间，视频总时长为 100 小时。视频被分割为 5秒的片段，并由 12名研究人员进行了标注。

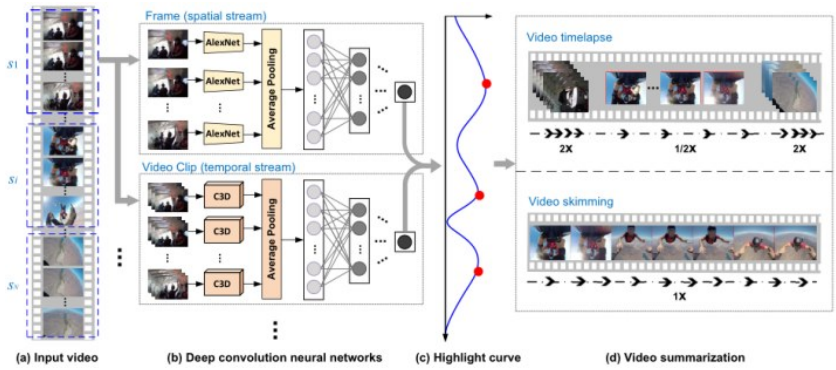

图 2　Yao 等提出的方法框架图

2.2.2　基于三维时空注意力网络的方法

Jiao 等[13]认为现有的大多数视频精彩集锦生成方法都是从整个视频片段中提取特征，而不考虑局部特征在时间和空间上的差异。在时间范围上，并非所有的帧都值得观看，而在空间范围上，并非每个帧的所有区域都是精彩的。为了解决上述问题，Jiao 等提出了一种新的三维时空注意力模型，该模型可以自动定位视频中的关键元素。具体地说，提出的注意模型沿着视频片段的空间和时间维度产生局部区域的注意权重。视频中关键元素的区域将通过大权重得到加强。因此，可以更有效的生成视频精彩集锦。Jiao 等提出的基于三维时空注意力模型的深度排序神经网络如图 3 所示，包括三个部分：特征模块、注意模块和排序模块。输入是一个原始视频片段。注意模块的功能是在空间和时间维度上同时选择重要的局部区域。然后排名模块预测最精彩片段的分数，获得视频的精彩程度曲线。

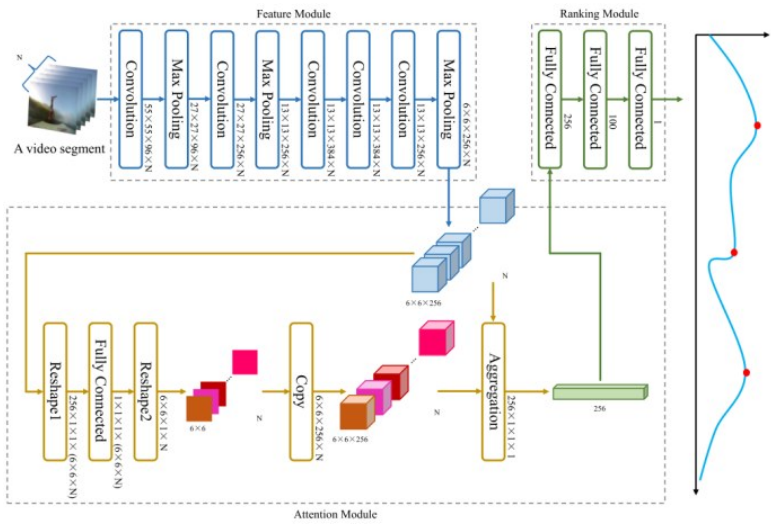

图 3　Jiao 等提出的深度排序神经网络流程图

3 无监督/弱监督学习方法

无监督学习训练样本的标记信息未知,目标是通过对无标记训练样本的学习来揭示数据的内在性质及规律,为进一步的数据分析提供基础。弱监督学习和无监督学习类似,但使用的训练数据的标注并不完全。

基于无监督/弱监督学习的视频精彩集锦生成技术通常是针对特定领域的,并基于公共性分析的思想,即在大量未标记的视觉数据中寻找低水平的视觉相关性或推断视频突出显示的公共特征。Chu等[14]发现,给定一组拥有同一主题的视频,重要的视觉概念往往会在不同的视频中反复出现。因此,视觉共现的频率被用来衡量视频片段的重要性。尽管这种方法的训练数据易于收集和拓展,但缺乏基本的真值标签使得学习一个有辨别力和鲁棒性的模型变得困难。基于无监督/弱监督学习的视频精彩集锦生成技术大都使用了深度学习方法。

3.1 基于鲁棒循环自动编码器的方法

Yang等[15]认为基于监督学习的方法依赖于成对的精彩集锦和原视频来推断视频的精彩部分。然而,想要收集这样的视频并不简单,用户通常不会同时上传视频的原始版本和编辑版本。为了解决这一问题,Yang等提出了一种无监督的视频精彩集锦生成方法,这一方法只使用人们编辑过的视频作为训练数据。Yang等设计了一个自动编码器,它有两个特点:一是使用了一种新的收缩指数损失函数,使自动编码器对噪声数据具有鲁棒性;另一个特点是编码器具有双向长短期记忆单元,以便在时间序列中有效地建模远程上下文。集锦生成算法的整体的架构如图4所示。每个视频首先被分割成多个短片段,然后应用预先训练好的3D卷积神经网络模型来提取时空特征,经过池化层后,使用设计的自动编码器来捕获远程上下文结构。Yang等从YouTube上收集了6500段短时长视频作为训练数据,这些数据没有进行额外标注。

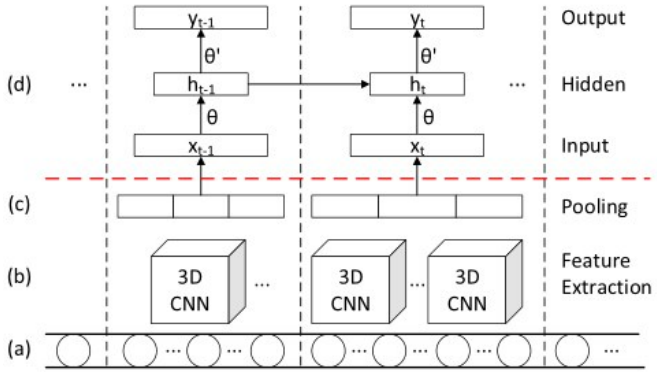

图4 Yang等提出的无监督方案架构图

3.2 使用时长作为隐式监督信号的方法

Xiong等[16]提出了一种可行的无监督解决方案,利用视频持续时间作为隐含的监督信号。Xiong等认为,用户生成视频中,较短时长视频的片段比较长时长视频的片段更有可能成为精彩集锦,因为用户在制作短时长视频时往往对内容进行了充分的选择。根据这一观点,Xiong等引入了一个新的排序框架,该框架优先选择短时长视频中的片段,同时适当考虑未标记的训练数据中的固有噪声。Xiong等还设计了一个新的损失函数,这个损失函数在长时长视频片段得分高时会增加。Xiong等在Instagram上收集了15种、超过1000万个视频用于训练,并在两个公共数据集TVSum[17]和YouTube Highlights[18]上进行了测试。图5展示了Xiong等收集的视频时长的分布。

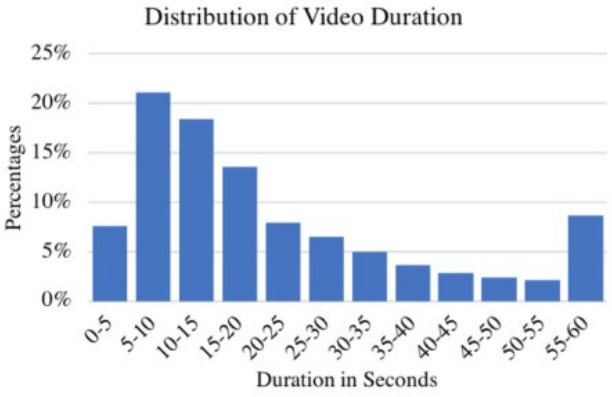

图5 Xiong等收集的视频时长分布

3.3　基于多流网络的无监督学习方法

Wang 等[19]研究了“王者荣耀“游戏视频的精彩镜头检测，使用没有额外注释的游戏视频作为训练数据，构造了一个包括时间流、空间流和音频流的多流网络。Wang 等下载了 450个经过剪辑的精彩集锦视频和 10个长时长的原始游戏视频，精彩集锦视频的平均长度为 21秒，而原始游戏视频的长度为 6 到 8 小时。由于原始游戏视频的长度非常长，Wang 等从视频中随机截取了 20个视频片段，每个视频平均长度为 13分钟，以平衡正负样本。Wang 等构建的多流网络结构如图 6 所示。该多流网络结合了三个组件来生成视频精彩集锦：时间流提取时态信息，使用三维卷积层[20]从 ResNet-34[21]最后的池化层的输出中提取特征；空间流获取每一帧的空间上下文信息，和时间流不同，空间流在帧级别上提取特征，使用了 AlexNet[22]；音频流通过利用声音特征过滤无关场景，使用了一个预训练的扬声器编码器。得到三个流输出的分数后，通过加权求和形成最终的分数，时间流、空间流和音频流分数的权重分别为 0.7、0.15、0.15，这表明了 3D 信息的重要性。

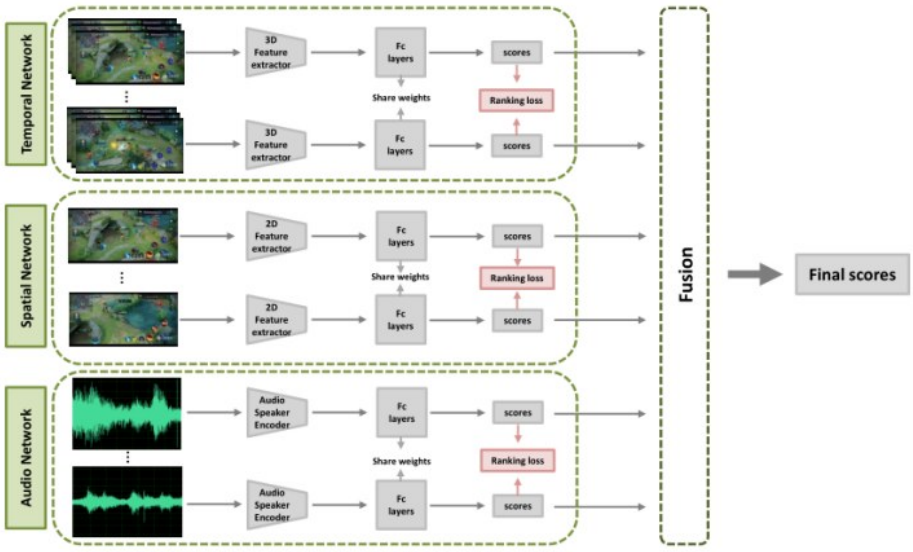

图 6　Wang 等构建的多流网络结构

3.4　基于实时评论的视频精彩集锦生成方法

近年来，互联网上开始流行实时评论，在弹幕平台网站上，观众可以在屏幕上发送实时评论（弹幕）来分享他们对视频的感受。实时弹幕与该时刻视频内容高度相关，实时评论是观众情绪的表达或对视频的讨论，视频越吸引观众，观众发布的实时评论就越多。因此，实时评论的数量在某种程度上可以反映这段视频的受欢迎程度。基于此，Wang 等[23]提出了基于实时评论生成视频精彩集锦的模型，该模型使用了卷积神经网络（CNN）和长短期记忆网络（LSTM），利用实时评论作为先验知识来辅助视频内容的分析，可以预测视频的精彩部分以及观众观看视频时的情绪。这一模型包括了两个子模块，分别为视频编码器和语言转换模块，视频编码器模块将视频序列编码为特征向量，语言转换模块将视频内容转换为人类语言的语义向量。

3.5　使用情感知识驱动的精彩视频集锦生成方法

视频精彩集锦生成是根据用户的兴趣选择一部分帧。Qi 等[24]认为传统的有监督学习方法的性能高度依赖于大规模人工标注的训练数据，这些数据的收集既耗时又费力。为了解决这个问题，Qi 等发现用户是否对特定的视频片段感兴趣在很大程度上取决于人类的主观情绪。利用这一观点，Qi 等设计了一个情感知识驱动的视频精彩集锦生成方法，用于建模人类的一般情感和推断视频的精彩程度。其设计的方法框架如图 7 所示。首先，通过前端网络获得视频片段的概念级表示，这些概念被用作构建情绪相关知识图的节点，它们在图中的关系通过外部公共知识图建

模。然后使用孪生图神经网络(Siamese GCN)对图中节点之间的依赖关系进行建模,并沿边传播消息。图神经网络能够转移视频上下文中出现的视觉概念的先验知识,以理解视频的高级语义。最后基于图神经网络层计算视频片段的情感感知表示,并进一步使用它预测精彩程度分数。

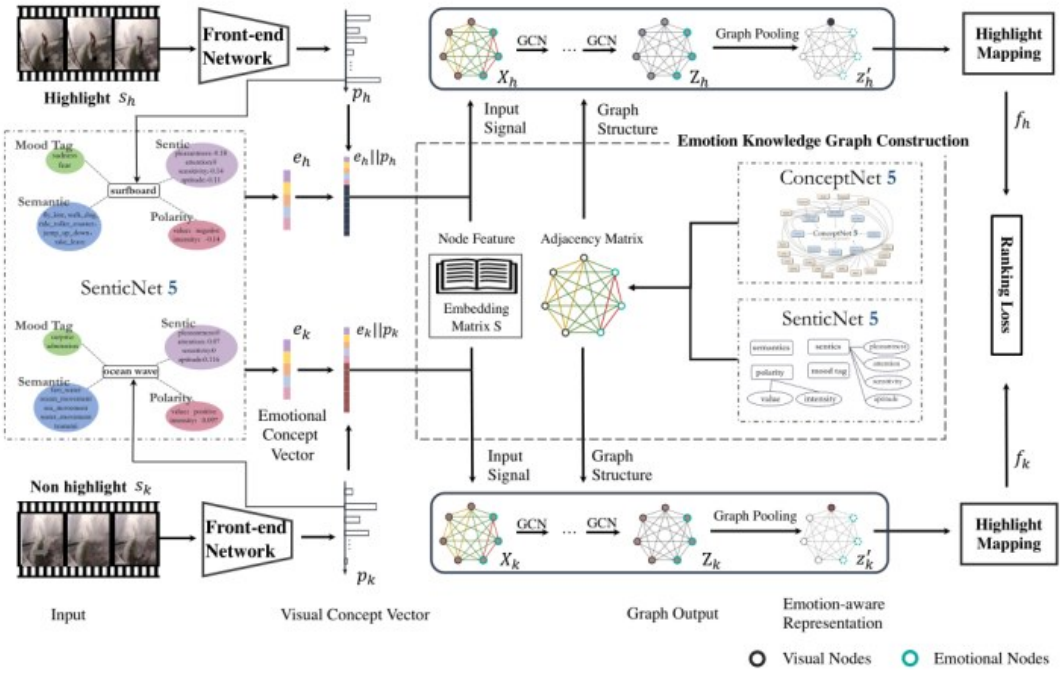

图7 Qi等构建的多流网络结构

4 技术应用

视频精彩集锦生成技术的目的是自动选取视频最有吸引力的片段,由于人工对视频进行剪辑需要大量的时间和精力,而现实生活中视频剪辑有着巨大的需求量,因此视频精彩集锦生成技术有很大的实用价值且在现实生活中有许多应用场景:

(1)对视频制作者而言,视频精彩集锦生成技术可以帮助他们自动对视频进行剪辑,生成的精彩集锦可以更好地吸引人们的兴趣,增加视频的播放量。无论是业余爱好者还是专业的视频制作者,视频精彩集锦生成技术都可以帮助他们减少人工剪辑视频的工作量。

(2)对视频网站而言,视频精彩集锦生成技术可以帮助网站吸引用户兴趣。相比于用视频封面吸引用户点击,使用视频精彩集锦生成技术自动生成的时长较短的集锦作为视频封面可以更好的吸引用户,促使他们观看完整的视频。

(3)视频精彩集锦生成技术还可以应用到电子商务平台的视频推荐系统。在电子商务中,产品相关视频是介绍产品特征、吸引消费者的重要内容。因此在电子商务平台的推荐系统中,可以使用视频精彩集锦生成技术来生成最具吸引力的视频片段展示给消费者以提高产品的点击率。例如,Guo等[25]提出了一种基于图形的商品感知模型,解决了电子商务场景中的多模态视频精彩集锦检测问题。

5 总结

本文调研了视频精彩集锦生成技术近年来的研究和发展情况。首先介绍了视频精彩集锦生成任务的定义,而后系统地梳理了相关的研究工作,介绍了视频精彩集锦生成技术的有监督学习方法和无监督/弱监督学习方法,并分析这两类方法的优缺点,最后介绍了视频精彩集锦生成技术在现实生活中的应用价值。

参考文献(References):

[1] Rui Y, Gupta A, Acero A. Automatically extracting highlights for TV baseball programs[C]. Proceedings of the Eighth ACM International Conference on Multimedia,.

2000: 105-115.

[2] Wang J, Xu C, Chng E, et al. Sports highlight detection from keyword sequences using HMM[C]. IEEE International Conference on Multimedia and Expo (ICME), 2004, 1: 599-602..

[3] Xiong Z, Radhakrishnan R, Divakaran A, et al. Highlights extraction from sports video based on an audio-visual marker detection framework[C]. IEEE International Conference on Multimedia and Expo, 2005: 4.

[4] Tang H, Kwatra V, Sargin M E, et al. Detecting highlights in sports videos: Cricket as a test case[C]. IEEE International Conference on Multimedia and Expo, 2011: 1-6.

[5] Sun M, Farhadi A, Seitz S. Ranking domain-specific highlights by analyzing edited videos[C].European Conference on Computer Vision, 2014: 787-802.

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

[7] Gygli M, Song Y, Cao L. Video2gif: Automatic generation of animated gifs from video[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1001-1009.

[8] Panda R, Das A, Wu Z, et al. Weakly supervised summarization of web videos[C].Proceedings of the IEEE International Conference on Computer Vision, 2017: 3657-3666.

[9] Potapov D, Douze M, Harchaoui Z, et al. Category-specific video summarization[C].European Conference on Computer Vision, 2014: 540-555.

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

[11] Liu W, Mei T, Zhang Y, et al. Multi-task deep visual-semantic embedding for video thumbnail selection[C] Proceedings of the IEEE Conference on Computer. Vision and Pattern Recognition, 2015:3707-3715.

[12] Nepal S, Srinivasan U, Reynolds G. Automatic detection of'goal'segments in basketball videos[C]. Proceedings of the ninth ACM International Conference on Multimedia, 2001: 261-269.

[13] Jiao Y, Li Z, Huang S, et al. Three-dimensional attention-based deep ranking model for video highlight detection [J]. IEEE Transactions on Multimedia, 2018, 20(10): 2693-2705.

[14] Chu WS, Song Y, Jaimes A. Video co-summarization: Video summarization by visual co-occurrence[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015:3584-3592.

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

[16] Xiong B, Kalantidis Y, Ghadiyaram D, et al. Less is more: Learning highlight detection from video duration[C].Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 1258-1267.

[17] Song Y, Vallmitjana J, Stent A, et al. Tvsum: Summarizing web videos using titles[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5179-5187.

[18] Sun M, Farhadi A, Seitz S. Ranking domain-specific highlights by analyzing edited videos[C].European Conference on Computer Vision, 2014: 787-802.

[19] Wang L, Sun Z, Yao W, et al. Unsupervised Multi-stream Highlight detection for the Game" Honor of Kings"[DB/OL]. arXiv preprint arXiv:1910.06189, 2019.

[20] Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6546-6555.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[22] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[23] Wang Z, Zhou J, Ma J, et al. Discovering attractive segments in the user-generated video streams[J]. Information Processing & Management, 2020, 57(1): 102130.

[24] Qi F, Yang X, Xu C. Emotion knowledge driven video highlight detection [J]. IEEE Transactions on Multimedia, 2020, 23: 3999-4013.

[25] Guo Z, Zhao Z, Jin W, et al. TaoHighlight: commodity-aware multi-modal video highlight detection in e-commerce [J]. IEEE Transactions on Multimedia (Early Access), 2021.

编辑：王谦，王雨田