文章编号:1672-3961(2024)03-0001-11 DOI:10.6040/j.issn.1672-3961.0.2023.109

短视频场景分类方法综述

聂秀山[1],巩蕊[1],董飞[2],郭杰[1*],马玉玲[1]
(1.山东建筑大学计算机科学与技术学院, 山东 济南 250101; 2.山东师范大学新闻与传播学院, 山东 济南 250358)

摘要:传统的视频场景分类方法习惯于从视觉模态中提取表现图像场景的特征,结合支持向量机等有监督学习方法,实现对某些类别的场景分类。随着各种短视频在各大平台迅速涌现,基于短视频特性的场景特征表示越来越受到研究者们的关注。由于短视频数据具有噪声、数据缺失、各模态语义强度不一致等问题,导致传统的视频场景表征方法无法学习具有丰富语义的短视频场景表征。近年来,部分短视频场景分类的研究考虑上述挑战,并提出相应的方法。本研究综述短视频场景分类的研究现状,介绍短视频场景特征表示和分类方法,对不同数据集上的场景分类方法进行分析。针对现有方法存在的问题,分析未来短视频场景分类中需要解决的挑战性问题。

关键词:视频场景;特征表示;短视频场景分类;多模态融合;深度学习
中图分类号:TP391 **文献标志码**:A
引用格式:聂秀山,巩蕊,董飞,等. 短视频场景分类方法综述[J].山东大学学报(工学版),2024,54(3):1-11.
NIE Xiushan, GONG Rui, DONG Fei, et al. A survey of micro-video scene classification[J]. Journal of Shandong University (Engineering Science), 2024, 54(3):1-11.

A survey of micro-video scene classification

NIE Xiushan[1], GONG Rui[1], DONG Fei[2], GUO Jie[1*], MA Yuling[1]
(1. School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, Shandong, China; 2. School of Journalism and Communication, Shandong Normal University, Jinan 250358, Shandong, China)

Abstract: Traditional video scene classification methods were used to extract the features of image scenes from the visual modality, and combined with supervised learning methods such as support vector machine to achieve scene classification of certain categories. With the rapid emergence of various micro-videos on major platforms, the scene feature representation based on the characteristics of micro-videos had attracted more and more attention of researchers. Due to the problems of micro-video data such as noise, data loss, and inconsistent semantic intensity of each modality, these issues resulted in traditional methods for representing video scenes being unable to learn micro-video scene representations with rich semantics. In recent years, the research of some micro-video scene classification had considered the above challenges and proposed corresponding methods based on micro-video scene classification. This study reviewed the research status of micro-video scene classification, introduced the feature representation and classification methods of micro-video scene, and analyzed the scene classification methods on different datasets. Aiming at the problems existing in the existing methods, the challenging problems to be solved in the future micro-video scene classification were analyzed.

Keywords: video scene; feature representation; micro-video scene classification; multi-modality fusion; deep learning

0 引言

场景是指从观察者的视角看所能观察或者活动的空间[1]。视频场景分类是对视频中包含的静态或者动态场景进行分类,是视频内容理解的重要任务。常见的场景大致可以分为几类:自然场景、城市场景、室内场景、户外场景和事件场景等[2-5]。近年来,随着计算机视觉的快速发展,涌现出众多视频场景分类的方法。传统的视频场景分类方法

收稿日期:2023-05-29
基金项目:国家自然科学基金资助项目(62176141,62176139,61876098);山东省杰出青年自然科学基金资助项目(ZR2021JQ26);山东省自然科学基金资助项目(ZR2021QF119)
第一作者简介:聂秀山(1981—),男,江苏徐州人,教授,博士生导师,博士,主要研究方向为机器学习与数据挖掘、视觉数据智能检索与分析。E-mail:niexiushan@163.com
*通信作者简介:郭杰(1990—),女,山东济南人,讲师,硕士生导师,博士,主要研究方向为多模态学习与智能多媒体计算。E-mail:guojiesdu@163.com

习惯于从视觉模态中提取表现图像场景的特征，例如通过提取全局特征（GIST）信息[6-7]、尺度不变特征变换（scale invariant feature transform，SIFT）[8]、加速稳健特征（speeded up robust feature，SURF）[9]、方向梯度直方图特征（histogram of oriented gradient，HOG）[10]、局部特征 GENTRIST[11-12]等图像的底层全局或局部的颜色、纹理、形状等特征，结合支持向量机等有监督学习方法实现对视频场景的分类。GIST 是图像的全局特征，通过图像的光谱信息反映全局信息。SIFT、HOG、GENTRIST 等是图像的局部特征，通过统计图像像素块的梯度方向信息反映其局部结构。GIST 特征计算复杂度低且使用较简单，能够反映场景的整体布局，但在背景复杂的场景中表现较差，因此比较适合在一些简单的自然场景分类任务中应用。SIFT 特征能够考虑到场景中的目标位置信息，对平移缩放、尺度变换、亮度变换、视角变化都有一定的稳定性，比较适合应用在大部分自然场景和一些简单的室内场景分类任务中。HOG 特征能够捕捉到场景的几何结构和整体布局，描述场景的边缘形状和目标轮廓，适用于场景结构形状稳定的情况。GENTRIST 特征通过图像的局部结构信息反映它的整体结构，对缩放和旋转具有不变性，适用于有清晰布局的场景分类任务。

由于视频包含时序信息[13-14]，因此时空特征表示成为视频场景特征表示的关键。研究人员提出多种方法表征视频场景的时空特征，包括时序特征与空间特征分别建模及直接对时空特征建模的方式。对于时空分开建模的方式，部分工作采用线性动态系统对场景在时间轴上的变化规律进行建模[15]，采用混沌理论将混沌不变量与全局静态特征相融合[16]，采用光流法[17]或者光流与空间金字塔的结合[18]对时序特征进行建模。对于直接时空特征建模的方式，部分研究提出时空方向能量特征包[19-20]、与多尺度相结合的时空方向能量特征[21]、通过空间场景块[22]表征场景的时空特征。还有研究通过提取视频中的慢特征[23-24]表征视频场景。随着深度学习框架在计算机视觉领域的发展，卷积神经网络也应用于视频场景分类中。三维卷积[25]及长短特征表示[26]方法也用于视频场景的时空特征表示。

自 2016 年以来，各种短视频社交媒体平台迅速涌现。与传统视频不同，短视频的生成基于用户的高度主观性，这为短视频场景分类带来以下挑战：（1）短视频场景的多模态之间一致性和互补性的协同问题；（2）噪声信息多；（3）同一场景的数据类内紧致性差；（4）部分模态数据缺失；（5）各模态语义强度不一致；（6）数据类别不平衡等。这些问题导致传统的视频场景表征方法无法学习具有丰富语义的短视频场景表征。近年来，部分短视频场景分类的研究考虑了上述挑战，并提出了相应的方法。因此，本研究对现有的短视频场景分类方面的研究进行综述，并对国内外研究现状进行分析。

1 短视频场景特征表示

在短视频场景分类问题中，其网络结构在逻辑上分为特征提取和分类 2 个阶段。视觉特征所传达的场景或视觉概念是场景类别的直观信号，从视觉模态中提取高级语义表示短视频尤为重要。例如从一个短视频中观察到桌子、椅子、咖啡杯等，能够很容易地预测出这个短视频是在咖啡厅拍摄的，这样能够从视觉模态中提取丰富的特征表示短视频。深度卷积神经网络（convolutional neural network，CNN）已作为提取视觉特征表示的强大模型。现有的短视频场景特征提取方法大多采用深度卷积神经网络提取场景的视觉特征，主要模型包括基于 AlexNet 与长短期记忆网络（long short-term memory，LSTM）结合的方法、基于 VGG16 与 Attention 结合的方法、基于 ResNet 与 Visual Transformer 的方法等。

1.1 基于 AlexNet 与 LSTM 结合的方法

2012 年的 ImageNet 大规模视觉识别挑战赛上提出了 AlexNet[27]并获得优异成绩，推动了卷积神经网络的发展。AlexNet 共由 5 个卷积层-最大池化层和 3 个全连接层组成，前 5 个卷积层-最大池化层用作特征提取，后 3 个全连接层作为分类器。拥有 6 000 万个参数和 650 000 个神经元。为了让训练更快，使用非饱和神经元和图形处理器实现。网络结构如图 1 所示。

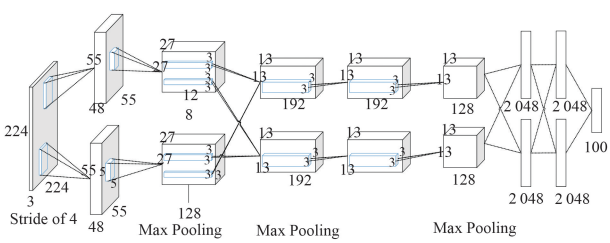


图 1 AlexNet 网络结构图

Fig.1 AlexNet network structure

AlexNet 中包含 Conv 卷积层，主要作用是提取输入图像的特征。提出局部响应归一化层，将数据归一化到 0~1，对局部神经元的活动创建竞争机制，使得响应较大的值变得相对更大，并抑制其他反馈较小的神经元，增强了模型的泛化能力。使用 ReLU 激活函数，成功解决了 Sigmoid 在网络较深时

的梯度弥散问题。MaxPool 为最大池化层,主要作用是对特征进行下采样,减小特征图的大小,同时保留特征的主要信息。AlexNet 中全部使用最大池化层,避免平均池化的模糊化效果,提升了特征的丰富性。

AlexNet 模型在 ILSVRC12 上的 120 万张干净图像上进行了预训练,可以为识别语义提供鲁棒的初始化。在提取特征之前,对每个短视频进行关键帧提取,对视频的所有关键帧采取平均池化策略,然后使用 AlexNet 提取每一帧的视觉特征,能够更好地提取视觉模态中的高级语义信息,并捕获视频帧之间的关系信息,有利于短视频场景的分类。

文献[28-29]使用 AlexNet 提取视觉特征,在此基础上,结合 LSTM 可以独立捕获特征序列[30]。使用 LSTM 提取帧级特征,可以将视频的时间结构捕获到单个表示中。其网络结构如图 2 所示。

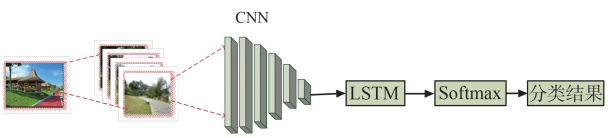


图 2　CNN-LSTM 网络结构图
Fig.2　CNN-LSTM network structure

LSTM 是一种循环神经网络(recurrent neural network, RNN)[31],是为了解决一般的循环神经网络存在的长期依赖导致的梯度消失或梯度爆炸的问题而设计的。LSTM 与一般的循环神经网络结构基本是一致的。前后连接的多个模块 A,同样是隐藏层在时间维度上的展开。但其单个神经元的结构与一般的循环神经网络稍有不同,增加了 3 个门控操作,分别为输入门 $\boldsymbol{i}_t$、遗忘门 $\boldsymbol{f}_t$ 和输出门 $\boldsymbol{o}_t$。这 3 个门控操作起到特征选择的作用。除此之外,隐藏层单元增加了 1 个输出,叫做细胞状态 C_t。

长短时记忆网络中各变量的计算公式为:

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f),$$
$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i),$$
$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o),$$
$$\check{C}_t=\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c),$$
$$C_t=\boldsymbol{f}_t\times C_{t-1}+\boldsymbol{i}_t\times\check{C}_t,$$
$$h_t=\boldsymbol{o}_t\times\tanh(C_t),$$

式中:$\boldsymbol{h}_{t-1}$为前一时刻的隐藏状态输出,其与当前时刻的输入 $\boldsymbol{x}_t$ 一起,作为新的输入;σ 是激活函数;$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是可学习的参数。

1.2　基于 VGG16 与 Attention 结合的方法

VGG 网络[32-33]在 2014 年提出,因其网络结构相对简单且网络训练过程中有着优异的表现,成为比较受欢迎的卷积神经网络模型。VGG16 包含 13 个卷积层、3 个全连接层、5 个池化层和一个 Softmax 层。网络结构如图 3 所示。

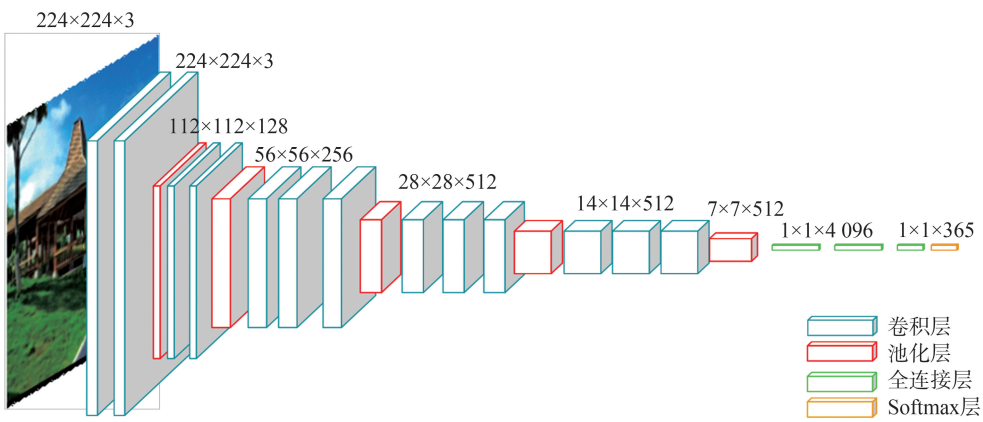


图 3　VGG16 网络结构图
Fig.3　VGG16 network structure

VGG16_places365 是 1 个预训练的 VGG16 网络,用于图像场景识别的公共数据集 Places365 中的图像识别。VGG16 网络由小卷积核、小池化核、ReLU 构成,结构相对简洁,通过增加深度能有效地提升性能,卷积可以代替全连接,可以适应各种尺寸的图片。使用 VGG16_places365 网络提取原始视觉特征,简化了卷积神经网络的结构,提高了训练的拟合能力,能够更好地保留高级语义信息。VGG16 在提取特征的过程中,使用大小为 3×3 的卷积核,使得训练结果更好。在卷积操作的过程中,中间位置的数值会多次提取,而边界数值的特征提取次数相对较少,加入了 padding 操作,能够更好地利用边界数值,也更方便计算。在卷积操作后提取的特征信息可能会存在信息冗余,通过池化层不断地减少数据空间的大小,使参数的数量和计算量不断地下降,能在一定程度上控制过拟合。

文献[34-35]使用 VGG16_places365 提取视觉特征,在此基础上,结合自注意力机制方法可以关注到更有效的特征信息,其网络结构如图 4 所示。自注意力机制在计算能力有限的情况下,将计算资源分配给更重要的任务,同时解决信息超载问题。在特征提取阶段,通过扫描全局图像,获取需要重点关注的目标区域,然后对这一区域投入更多的注意力,获取更多有价值的细节信息,提取更丰富的语义信息。

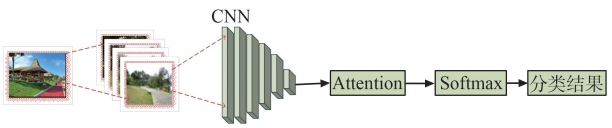


图 4 CNN-Attention 网络结构图
Fig.4 CNN-Attention network structure

1.3 基于 ResNet 与 Visual Transformer 结合的方法

在短视频视觉特征提取过程中,研究人员发现,随着神经网络层数的增加,网络性能越来越低,显示出退化问题。为了解决这一问题,2015 年的 ImageNet 比赛提出 ResNet 并取得优异成绩[36],ResNet 设计一种使用 skipconnection 的残差结构,使得网络达到很深的层次,同时提升了性能。残差结构如图 5 所示。以 ResNet50 为例,共包含 49 个卷积层、1 个全连接层。网络结构可以分成 7 部分,第 1 部分主要对输入进行卷积、正则化、激活函数、最大池化的计算,第 2、3、4、5 部分包含了残差块,每个残差块包含 3 层卷积。网络的输入为 224×224×3,经过前 5 部分卷积计算,输出为 7×7×2 048,池化层会将其转化成 1 个特征向量,最后分类器会对这个特征向量进行计算并输出。

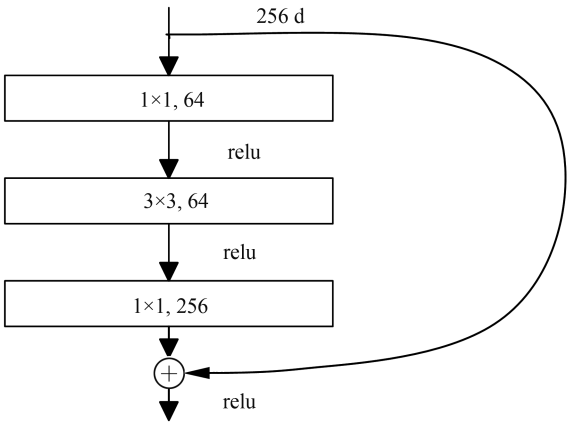


图 5 ResNet 残差结构图
Fig.5 ResNet residual structure

ResNet 在多个领域都有广泛应用,在场景分类任务中,ResNet 网络能够捕捉到场景类别的突出视觉特征,使得视觉特征传递出丰富的语义信息。因其可以训练非常深的神经网络,能够避免梯度消失问题,从而提高模型的表达能力和性能。通过使用残差结构保留原始特征,使得网络的学习更加顺畅和稳定,进一步提高了模型的精度和泛化能力。在训练时可以避免梯度消失和梯度爆炸问题,加速网络收敛,从而使场景分类方法达到较高的性能。

文献[37]使用 ResNet 提取视觉特征,在此基础上,文献[38]提出了域自适应网络 ResNet-DT,有效地提高特征的质量。为了更有效聚焦短视频的场景信息,文献[39]提出了 VT-ResNet 提取视觉特征,取得了较好的效果,其网络结构如图 6 所示。

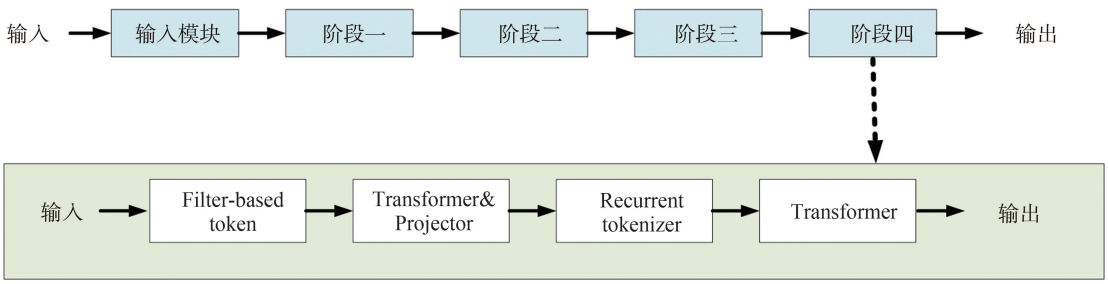


图 6 VT-ResNet 网络结构图
Fig.6 VT-ResNet network structure

对于声音特征,大部分研究者通过去噪自编码器(denoising autoencoders, DAE)[40-41]提取声音特征,其原理如图 7 所示。

$\boldsymbol{x}$ 作为输入被 q_D 随机破坏到 $\boldsymbol{x}_D$,自编码器 f_m 将其映射到 $\boldsymbol{y}$,并通过解码器 g_m 重建 $\boldsymbol{x}$,产生重建 $\boldsymbol{z}$,重建误差通过 $L_H(\boldsymbol{x},\boldsymbol{z})$ 计算。文本特征一般通过 Sentence2vector[42-43]、word2vector[44]、Paragraph Vector[45] 等提取,Paragraph Vector 网络结构如图 8 所示。

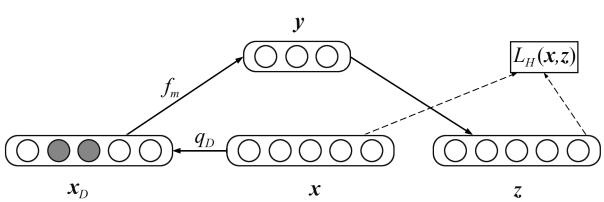


图 7 降噪自编码器结构图
Fig.7 Structure of noise reduction autoencoder

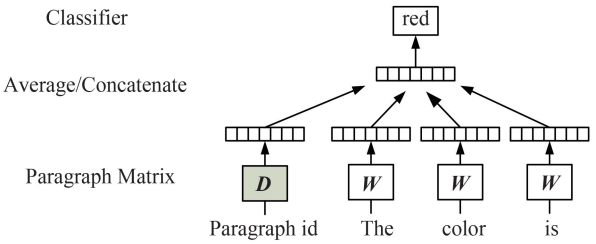


图 8　Paragraph Vector 提取器结构图
Fig.8　Paragraph Vector extractor structure

2　短视频场景分类方法

相比传统的视频场景分类方法,短视频场景分类大多基于短视频特性学习场景特征表示,因此可以较好地表征短视频场景语义。由于短视频中包含视频、声音、评论文本等多模态的信息,因此现有的短视频场景分类方法大多基于多模态融合。还有部分工作仅在视觉模态中基于短视频特性研究场景特征表示。以下从单模态短视频场景分类和基于多模态融合的短视频场景分类 2 个方面对现有工作展开综述和分析。

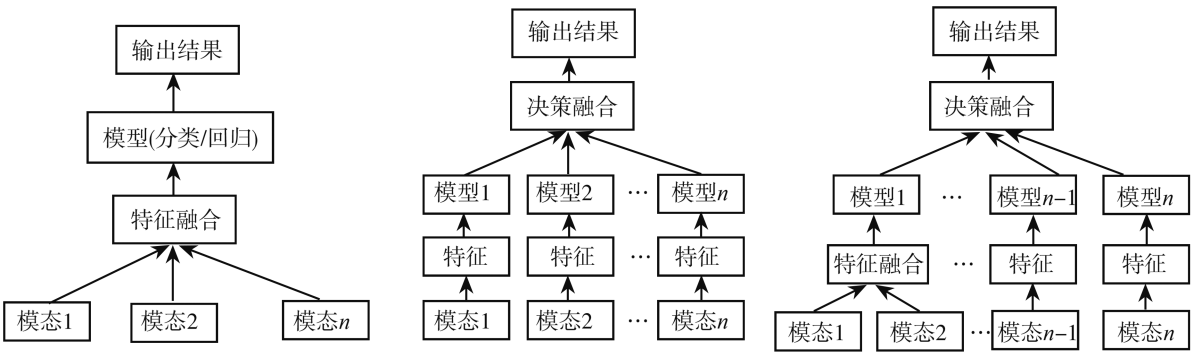
2.1　单模态短视频场景分类

虽然多模态的信息表示可以为短视频场景提供更丰富的语义特征,但由于视觉模态含有丰富的场景语义信息,因此部分研究基于短视频的特性对视觉模态的短视频场景表征开展。

考虑用户拍摄视频时的主观性对视频内容的影响,导致同一场景短视频存在视觉内容差异较大的问题,文献[46]提出基于注意力机制的一致性语义学习模型,该模型通过双分支注意力网络增强类内样本的语义一致性,该方法适用于短视频各个模态。为了增强视觉模态特征的语义性,文献[38]提出分层注意力和帧差增强网络。该网络包含运动和内容 2 个并行分支,首先,通过域自适应的卷积神经网络模型和时间-位移模型提取判别性的视觉特征,并通过分层注意力和 LSTM 增强时空特征;然后,通过 3D 卷积神经网络和长短期记忆网络提取视频片段中所包含的运动信息;最后,2 个分支融合作为短视频视觉特征表示。

2.2　基于多模态融合的短视频场景分类

基于模态融合的层级划分,可以将多模态融合策略分为早期融合[47-49]、晚期融合[50-51]和中期融合[52-53],如图 9 所示。



(a) 早期融合　　(b) 晚期融合　　(c) 中期融合

图 9　多模态融合策略示意图
Fig.9　Multi-mode fusion strategy

早期融合也称为特征级融合,是指特征层面和数据层面的融合,融合特征作为输入数据输入到一个模型中,输出预测结果;晚期融合也称决策级融合,先用不同模型对不同模态进行训练,再融合多个模型输出的结果;随着深度学习的兴起,神经网络可以进行特征的自动提取,利用网络的中间隐层对不同模态的特征进行融合,这种融合方式有别于在数据层面和决策层面的融合,因而称为中期融合。

现有基于多模态融合的短视频场景分类的研究大多聚焦于如何学习多模态之间的一致性和互补性,如表 1 所示。

表 1　基于短视频特性的场景特征表示方法
Table 1　Venue feature representation method based on micro-video feature

方法	特征	融合策略
ACSL[16]	视觉	
HAFDN[17]	视觉	
TRUMANN[18]	视觉、声音和文本	中期融合
Deep transfermodel[19]	视觉、声音和文本	早期融合
EASTERN[20]	视觉、声音和文本	中期融合
INTIMATE[21]	视觉、声音和文本	早期融合
Multi-layerneural network[23]	视觉、声音和文本	早期融合

表 1(续)

方法	特征	融合策略
Combinational fusion method[24]	视觉、声音和文本	早期融合
MESL[25]	视觉、声音和文本	晚期融合
Jointly learning model[26]	视觉、声音和文本	中期融合
NNeXtVLAD[27]	视觉、声音和文本	中期融合
Multi-modality sequence model[28]	视觉、声音和文本	早期融合
NMCL[29]	视觉、声音和文本	晚期融合
AETML[30]	视觉、声音和文本	晚期融合
DMLRD[35]	视觉、声音和文本	晚期融合
CDGNN[36]	视觉、声音和文本	晚期融合

其中部分研究通过公共子空间学习的方式学习多模态之间的一致性,另一部分研究通过特征串联组合的方式学习多模态之间的一致性和互补性。

2.2.1 基于公共子空间学习的场景分类

为了学习多模态之间的一致性,部分研究通过公共子空间学习的方式进行场景分类。如图 10 所示。

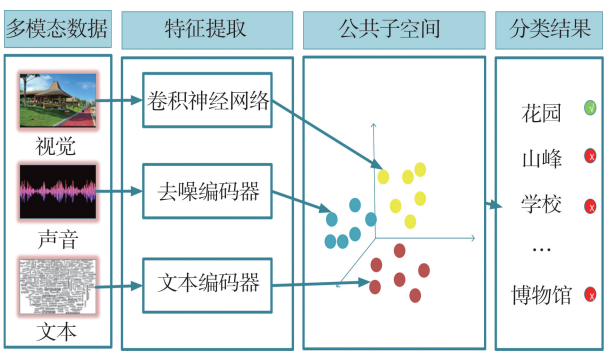


图 10 基于公共子空间学习的场景分类框架

Fig.10 Scene classification framework based on common subspace learning

针对短视频中的多模态信息及场景类别层次化分布的特点,文献[28]提出一种新的基于树引导的多任务多模态学习方法,标记具有场景类别的视频片段,该模型能够从不同模态中学习一个共同的特征空间,通过不一致惩罚保留每个模态的信息,从而得到更清晰的场景类别表示。针对短视频时长短、概念稀疏的特点,文献[54]提出一种端到端深度学习模型,该模型采用 3 个并行长短期记忆网络捕获序列结构,并采用卷积神经网络学习稀疏概念级表示,应用于短视频场景分类。考虑到训练样本的时效性和局限性,文献[55]又提出一种有效的在线学习算法提高学习性能,并提出一种基于结构引导的多模态字典学习框架,该框架将统一模型内的层次平滑性和结构一致性协同正则化,学习短视频的高层稀疏表示。文献[56]提出 LSTMs-CNN 联合学习框架,通过 3 个并行的长短期记忆网络分别提取 3 个模态的特征表示,将其映射到公共子空间,将 3 个模态的特征表示输入卷积神经网络,得到最后的场景特征表示。

2.2.2 基于特征串联学习的场景分类

公共子空间学习的方式可以较好地学习多模态之间的一致性,但是不能很好地达到多模态之间一致性与互补性的平衡。为了更好地学习多模态之间的一致性和互补性,部分研究通过特征串联的方法进行场景分类,如图 11 所示。

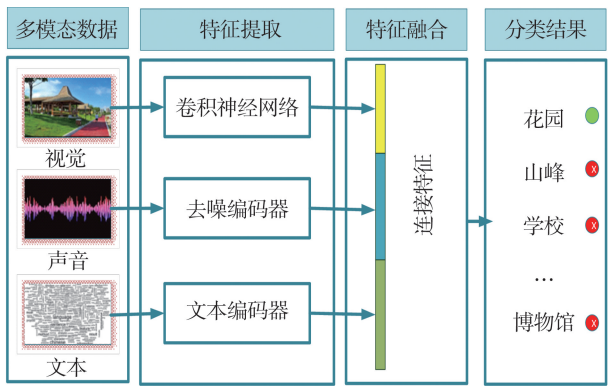


图 11 基于特征串联学习的场景分类示意图

Fig.11 Scene classification based on feature tandem learning

文献[57-58]提出带有门控全卷积模块的多模态序列模型,该模型通过门控全卷积模块分别提取 3 个模态的特征,然后连接起来通过门控全卷积块得到多模态融合后的特征表示。由于短视频中声音模态存在噪声信息,为了使声音模态在场景分类任务中发挥更大的作用,文献[29]通过利用外部声音知识,增强声音模态估计短视频的场景类别。为了更好地区分视频中的场景类别,文献[59]构建了 1 个语义框架,该框架基于潜在对象的视觉语义表示学习场景分类器,然后提出使用场景级和潜在对象级 2 种不同类型的迭代执行主动学习,最后采用 1 种自适应策略自动执行这 2 种类型的主动学习迭代之间的切换,从而实现一种新的多级自适应主动学习方法。

为了充分利用多模态之间的一致性和互补性,文献[34]将多个模态通过早期融合输入神经网络进行非线性变换,然后将最后 1 个隐藏层的输出作为判别场景特征,实现基于多模态融合的短视频场景分类。在此基础上,文献[35]提出一种联合多层神经网络和监督哈希学习的统一框架,采用多层神经网络通过非线性变换融合多种模态,通过监督哈希学习方法将融合特征转换为二进制代码,以保持语义和相似性。针对多模态语义强弱不一致的问

题,文献[60]提出一种用于场景识别的深度多模态融合网络。视觉模态认为是主要模态,其他模态(如文本和声音)认为是辅助模态。通过主模态对辅助模态的语义增强实现自适应权重的多模态融合。为了解决多个输出任务时表现出的局限性,文献[61]提出一种用于多标签短视频分类任务的双多模态低秩分解方法,学习更全面的短视频表示。同时,文献[62]还提出一种具有串行自注意机制的多模态聚合网络执行微视频多标签分类任务。

为了从多模态信息中明确分离出一致性特征和互补性特征,并利用它们的组合提高各模态的表达能力,文献[37]提出一种神经多模态合作学习模型,通过一种新的关系感知注意机制分割一致性成分和互补性成分。针对短视频各模态数据存在不确定性的问题,文献[39]提出一种注意力增强和可靠的多模态学习模型,通过构建一个不确定性估计网络动态评估此类决策的可靠性,以实现短视频场景分类的合理多模态决策融合,得出可靠的预测结果。

相比传统视频场景分类,上述研究大多基于短视频的特性学习场景分类方法,包括噪声、概念稀疏、模态间语义强弱不一致、同一场景类内紧致性差、数据缺失等问题。但基本采用的是比较简单的均值填充、强模态对弱模态进行语义增强等方法。没有考虑各模态噪声的分布、强弱模态自适应辅助增强等,不能充分表示原始语义信息。

3　数据集和性能度量

本章介绍了比较常用的视频场景分类的数据集、各方法的性能度量标准以及各数据集上现有方法的性能分析。

3.1　数据集

Vine[28]:该数据集为从 Vine 平台收集的真实数据以及通过其公共 API 从 Vine 捕获的短视频。数据集只包含 3 种模态、位置信息和恰好 6 s 的短视频。此外,该数据集包含 270 145 个短视频,188 个场景类别,每个类别的样本数量是不平衡的。

MicrovideoSceneData_10[46]:通过从 Vine 上下载并重组创建的新的短视频场景数据集。数据集包含 10 个场景类别,每个类别包含 100 到 2 000 个样本,每个视频样本由一个移动的摄像机拍摄,平均约 6 s。

Maryland dataset[20]:该数据集中的视频从 YouTube 中收集,共有 13 个动态场景类,每个类包含 10 个彩色视频,视频的平均尺寸为 308 像素×417 像素×617 帧。

YUPENN dataset[26]:该数据集包含与动态场景相关的 14 个不同的视频目录,每个目录中有 30 个样本。这些视频样本是用固定摄像机拍摄的,每个视频时长约 5 s,共 150 帧。

UCF-101 dataset[26]:该数据集是一个现实动作视频的动作识别数据集,共有 13 320 个视频,包含 101 个类别。使用 2/3 个样本进行训练,1/3 个样本进行测试。该数据集总时长大约 27 h,主要包括 5 大类动作:人与物体交互、单纯的肢体动作、人与人交互、演奏乐器、体育运动。每个类别分为 25 组,每组 4~7 个短视频。

YouTube-8M[63]:该数据集为 Google 在 2016 年发布的大规模视频数据集,包含 800 万个视频,这些数据集进行了视频层级的标注,每个视频至少有 1 000 帧,每个视频的长度为 120~500 s,视频总时长约为 50 万 h,并用 4 800 个视觉实体的词汇表进行注释。

MTSVRC dataset[61]:该数据集在 PRCV2018 大赛上由美图公司公开提供,由 10 万个短视频组成,其中训练集有 50 000 个短视频,验证集和测试集分别有 25 000 个短视频,包含画画、唱歌、健身、羽毛球等 50 个热门类别,除了包含与人有关的一些行为类别,还有一些风景、宠物等类别。每个短视频时长不超过 15 s。

3.2　性能度量标准

现有研究大部分用准确率 A、Micro-F_1(F_1_i)、Macro-F_1(F_1_a)、平均准确率 M 等表示方法的性能。准确率为预测正确样本在所有样本中所占的比例,适用于平衡的数据集。F_1_i 为所有样本分配相同的权重,适用于多分类不平衡的数据集;F_1_a 为每个类别分配相同的权重,不受数据不平衡的影响。平均准确率属于检索任务的通用标准。

3.3　方法性能分析

本研究对不同数据集上的方法进行了分析,通过性能标准指标评价了方法的性能。结果如表 2 所示。

由表 2 可以看出,Vine 数据集中各个方法的分类性能普遍较低,F_1_i 不高于 70%,F_1_a 不高于 50%。这是由于 Vine 数据集中的数据符合真实的数据分布,数据量较大,存在噪声问题;部分数据不完整,存在数据缺失,导致各模态语义强弱不一致;同时该数据集存在类内紧致性较差,数据类别不平

衡等问题。因此该短视频场景分类任务具有较大挑战性。从MicrovideoSceneData_10数据集中可以看出,多模态融合的方法相比单一模态的ACSL方法性能高。短视频场景的多模态之间存在一致性和互补性的协同问题,利用多模态的信息学习不同模态之间的一致性与互补性,达到不同模态之间一致性与互补性的平衡,学习更丰富的语义表征。因此,基于多模态融合的短视频场景分类具有非常大的研究空间。

表2 不同数据集上的方法性能分析

Table 2 Method analysis on different datasets

单位:%

数据集	方法	性能度量指标	性能值
Vine	TRUMANN[28]	F_1_i	25.27
		F_1_a	5.21
	Deep transfermodel[29]	F_1_i	31.21
		F_1_a	16.66
	NMCL[37]	F_1_i	40.04
		F_1_a	26.78
	EASTERN[54]	F_1_i	59.51
		F_1_a	30.57
	Jointly learningmodel[56]	F_1_i	62.73
		F_1_a	32.93
	Multi-modality sequencemodel[58]	F_1_i	63.23
		F_1_a	33.84
	NNeXtVLAD[57]	F_1_i	**66.87**
		F_1_a	**41.88**
	Multi-layer neuralnetwork[34]	$\boldsymbol{M}$(@50)	45.04
		$\boldsymbol{M}$(@100)	44.68
	Combinational fusionmethod[35]	$\boldsymbol{M}$(@50)	46.90
		$\boldsymbol{M}$(@100)	47.70
Microvideo Scene Data_10	MESL[60]	A	**98.26**
	ACSL[46]	A	75.30
	HAFDN[38]	F_1_i	86.60
		F_1_a	83.70
	AETML[39]	F_1_i	96.24
		F_1_a	96.38
public dataset	INTIMATE[55]	F_1_i	**6.60**
		A	6.28
Maryland	Christoph Feichtenhofer et al[20]	A	**77.69**
	LSTF[26]	A	95.00
YUPENN dataset	Arun Balajee Vasudevan et al[18]	A	85.61
	LSTF[26]	A	**98.00**
UCF-101 dataset	LSTF[26]	A	**85.20**
MTSVRC dataset	DMLRD[61]	M	82.18
	CDGNN[62]	M	**89.91**

4 结论

本研究综述了短视频场景分类的研究现状,介绍了短视频场景特征表示及分类方法,对不同数据集上的场景分类方法进行了分析。与传统视频相比,短视频场景面临许多挑战,包括短视频场景的多模态信息表示、短视频数据噪声信息多、相同场景的短视频类内紧致性差、部分模态数据缺失、各模态语义强度不一致等。虽然已有的工作已经考虑了其中的部分问题,但并未对根源性问题进行深入研究。

(1)噪声分布与不确定性建模。短视频中各模态的噪声信息导致短视频场景数据具有不确定性。研究噪声分布和数据的不确定性建模,可以增强分类结果的可靠性。

(2)缺失数据处理。短视频中存在部分模态数据缺失的问题,研究缺失数据的处理,可以充分利用多模态数据之间的关联,从而增强特征的语义表征能力。

(3)模态内和模态间数据的语义一致性分析。短视频模态内和模态间都存在语义紧致性较弱的特点,对模态内和模态间数据的语义一致性进行分析,可以增强表征建模的准确性。

因此,未来短视频场景分类还需要针对这些问题进行深入探索。

参考文献:

[1] OLIVA A, TORRALBA A. Modeling the shape of the scene: a holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[2] SUDDERTH E B, TORRALBA A, FREEMAN W T, et al. Learning hierarchical models of scenes, objects, and parts[C]//Tenth IEEE International Conference on Computer Vision (ICCV'05): Volume 1. Piscataway, USA: IEEE, 2005: 1331-1338.

[3] ZUO Zhen, WANG Gang, SHUAI Bing, et al. Exemplar based deep discriminative and shareable feature learning for scene image classification[J]. Pattern Recognition, 2015, 48(10): 3004-3015.

[4] SINGH V, GIRISH D, RALESCU A L. Image understanding-a brief review of scene classification and recognition[J]. MAICS, 2017: 85-91.

[5] XIAO J, HAYS J, EHINGER K A, et al. SUN database: large-scale scene recognition from abbey to zoo[C]//

Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2010.

[6] OLIVA A, TORRALBA A. Modeling the shape of the scene: a holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3):145-175.

[7] OLIVA A, TORRALBA A. Building the gist of a scene: the role of global image features in recognition[J]. Progress in Brain Research, 2006, 155: 23-36.

[8] BROWN M, SÜSSTRUNK S. Multi-spectral SIFT for scene category recognition [C]//IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2011: 177-184.

[9] BAY H, ESS A, TUYTELAARS T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding, 2008, 110(3): 346-359.

[10] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05). Piscataway, USA: IEEE, 2005, 1: 886-893.

[11] WU J, REHG J M. Centrist: a visual descriptor for scene categorization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 33(8): 1489-1501.

[12] ZABIH R, WOODFILL J. Non-parametric local transforms for computing visual correspondence[C]//Computer Vision: ECCV′94: third European Conference on Computer Vision Stockholm: Volume II 3. Berlin, German: Springer, 1994: 151-158.

[13] FEICHTENHOFER C, PINZ A, WILDES R P. Spacetime forests with complementary features for dynamic scene recognition[C]//British Machine Vision Conference. Berlin, German: Springer, 2013: 6.

[14] GANGOPADHYAY A, TRIPATHI S M, JINDAL I, et al. Dynamic scene classification using convolutional neural networks[C]//2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Piscataway, USA: IEEE, 2016: 1255-1259.

[15] DORETTO G, CHIUSO A, YING N W, et al. Dynamic textures[J]. International Journal of Computer Vision, 2003, 51: 91-109.

[16] SHROFF N, TURAGA P, CHELLAPPA R. Moving vistas: exploiting motion for describing scenes[C]// IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2010: 1911-1918.

[17] MARSZALEK M, LAPTEV I, SCHMID C. Actions in context[C]//IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2009: 2929-2936.

[18] VASUDEVAN A B, MURALIDHARAN S, CHINTAPALLI S P, et al. Dynamic scene classification using spatial and temporal cues[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway, USA: IEEE, 2013: 803-810.

[19] FEICHTENHOFER C, PINZ A, WILDES R P. Dynamic scene recognition with complementary spatiotemporal features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(12):2389-2401.

[20] FEICHTENHOFER C, PINZ A, WILDES R P. Bags of spacetime energies for dynamic scene recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2014: 2681-2688.

[21] DERPANIS K G, LECCE M, DANIILIDIS K, et al. Dynamic scene understanding: the role of orientation features in space and time in scene classification[C]// IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2012: 1306-1313.

[22] DU Liang, LING Haibin. Dynamic scene classification using redundant spatial scenelets[J]. IEEE Transactions on Cybernetics, 2015, 46(9): 2156-2165.

[23] THERIAULT C, THOME N, CORD M. Dynamic scene classification: learning motion descriptors with slow features analysis [C]//IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2013: 2603-2610.

[24] WISKOTT L, SEJNOWSKI T J. Slow feature analysis: unsupervised learning of invariances[J]. Neural Computation, 2002, 14(4): 715-770.

[25] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[26] HUANG Yuanjun, CAO Xianbin, WANG Qi, et al. Long-short-term features for dynamic scene classification [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(4): 1038-1047.

[27] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[28] ZHANG Jianglong, NIE Liqiang, WANG Xiang, et al. Shorter-is-better: venue category estimation from micro-video[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York, USA: ACM, 2016: 1415-1424.

[29] NIE Liqiang, WANG Xiang, ZHANG Jianglong, et al. Enhancing micro-video understanding by harnessing external sounds [C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM, 2017: 1192-1200.

[30] GRAVES A. Long short-term memory[J]. Supervised Sequence Labelling with Recurrent Neural Networks, 2012, 385: 37-45.

[31] LIPTON Z C, BERKOWITZ J, ELKAN C. A critical review of recurrent neural networks for sequence learning [EB/OL]. (2015-10-17)[2023-05-18]. https://arxiv.org/abs/1506.00019.

[32] ZHOU B, LAPEDRIZA A, KHOSLA A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 40(6): 1452-1464.

[33] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-05-10)[2023-05-18]. https://arxiv.org/abs/1409.1556.

[34] GUO Jie, NIE Xiushan, CUI Chaoran, et al. Getting more from one attractive scene: venue retrieval in micro-videos[C]//Advances in Multimedia Information Processing-PCM 2018: 19th Pacific-Rim Conference on Multimedia. Berlin, German: Springer, 2018: 721-733.

[35] GUO Jie, NIE Xiushan, JIAN Muwei, et al. Binary feature representation learning for scene retrieval in micro-video[J]. Multimedia Tools and Applications, 2019, 78: 24539-24552.

[36] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision & Pattern Recognition. Piscataway, USA: IEEE, 2016: 770-778.

[37] WEI Yinwei, WANG Xiang, GUAN Weili, et al. Neural multimodal cooperative learning toward micro-video understanding[J]. IEEE Transactions on Image Processing, 2019, 29: 1-14.

[38] WANG Bing, HUANG Xianglin, CAO Gang, et al. Hybrid-attention and frame difference enhanced network for micro-video venue recognition[J]. Journal of Intelligent & Fuzzy Systems, 2022, 43(3): 3337-3353.

[39] WANG Bing, HUANG Xianglin, CAO Gang, et al. Attention-enhanced and trusted multimodal learning for micro-video venue recognition[J]. Computers and Electrical Engineering, 2022, 102: 108127.

[40] EL-NOUBY A, IZACARD G, TOUVRON H, et al. Are large-scale datasets necessary for self-supervised pre-training? [EB/OL]. (2021-12-20)[2023-05-18]. https://arxiv.org/abs/2112.10740.

[41] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM, 2008: 1096-1103.

[42] KIROS R, ZHU Y, SALAKHUTDINOV R R, et al. Skip-thought vectors[J]. Advances in Neural Information Processing Systems, 2015, 28: 1-9.

[43] ARORA S, LIANG Y, MA T. A simple but tough-to-beat baseline for sentence embeddings[C]//International Conference on Learning Representations. New York, USA: ICML, 2017: 1-16.

[44] RONG X. Word2vec parameter learning explained[EB/OL]. (2016-06-05)[2023-05-18]. https://arxiv.org/abs/1411.2738.

[45] LE Q, MIKOLOV T. Distributed representations of sentences and documents[C]//International Conference on Machine Learning. New York, USA: ACM, 2014: 1188-1196.

[46] GUO Jie, NIE Xiushan, MA Yuling, et al. Attention based consistent semantic learning for micro-video scene recognition[J]. Information Sciences, 2021, 543: 504-516.

[47] FAN Weiquan, HE Zhiwei, XING Xiaofen, et al. Multi-modality depression detection via multi-scale temporal dilated cnns[C]//Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. New York, USA: ACM, 2019: 73-80.

[48] YIN Shi, LIANG Cong, DING Heyan, et al. A multi-modal hierarchical recurrent neural network for depression detection[C]//Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. [S.l.]: ACM, 2019: 65-71.

[49] RAY A, KUMAR S, REDDY R, et al. Multi-level attention network using text, audio and video for depression prediction[C]//Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop[S.l.]: ACM, 2019: 81-88.

[50] MENG Hongying, HUANG Di, WANG Heng, et al. Depression recognition based on dynamic facial and vocal expression features using partial least square regression[C]//Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge. New York, USA: ACM, 2013: 21-30.

[51] SAMAREH A, JIN Y, WANG Z, et al. Detect depression from communication: how computer vision, signal processing, and sentiment analysis join forces[J]. IISE

Transactions on Healthcare Systems Engineering, 2018, 8(3): 196-208.

[52] NIE Weizhi, YAN Yan, SONG Dan, et al. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition[J]. Multimedia Tools and Applications, 2021, 80: 16205-16214.

[53] VERMA S, WANG J, GE Z, et al. Deep-HOSeq: deep higher order sequence fusion for multimodal sentiment analysis[C]//2020 IEEE International Conference on Data Mining (ICDM). Piscataway, USA: IEEE, 2020: 561-570.

[54] LIU Meng, NIE Liqiang, WANG Meng, et al. Towards micro-video understanding by joint sequential-sparse modeling [C]//Proceedings of the 25th ACM International Conference on Multimedia. New York, USA: ACM, 2017: 970-978.

[55] LIU Meng, NIE Liqiang, WANG Xiang, et al. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning[J]. IEEE Transactions on Image Processing, 2018, 28(3): 1235-1247.

[56] LIU Wei, HUANG Xianglin, CAO Gang, et al. Joint learning of LSTMs-CNN and prototype for micro-video venue classification[C]//Advances in Multimedia Information Processing: PCM 2018: 19th Pacific-Rim Conference on Multimedia. Berlin, German: Springer, 2018: 705-715.

[57] LIU Wei, HUANG Xianglin, CAO Gang, et al. Joint learning of nnextvlad, cnn and context gating for micro-video venue classification[J]. IEEE Access, 2019, 7: 77091-77099.

[58] LIU Wei, HUANG Xianglin, CAO Gang, et al. Multi-modal sequence model with gated fully convolutional blocks for micro-video venue classification[J]. Multimedia Tools and Applications, 2020, 79(9/10): 6709-6726.

[59] LI Xin, GUO Yuhong. Multi-level adaptive active learning for scene classification[C]// European Conference on Computer Vision. Berlin, German: Springer, 2014: 234-249.

[60] GUO Jie, NIE Xiushan, YIN Yilong. Mutual complementarity: multi-modal enhancement semantic learning for micro-video scene recognition[J]. IEEE Access, 2020, 8: 29518-29524.

[61] LU Wei, LI Desheng, NIE Liqiang, et al. Learning dual low-rank representation for multi-label micro-video classification[J]. IEEE Transactions on Multimedia, 2023, 25: 77-89.

[62] LU Wei, LIN Jiaxin, JING Peiguang, et al. A multimodal aggregation network with serial self-attention mechanism for micro-video multi-label classification[J]. IEEE Signal Processing Letters, 2023, 30: 60-64.

[63] ABU-EL-HAIJA S, KOTHARI N, LEE J, et al. YouTube-8M: a large-scale video classification benchmark [EB/OL]. (2016-09-27)[2023-05-18]. https://arxiv.org/abs/1609.08675.

(编辑:郭少华)