基于循环区域关注和视频帧关注的视频行为识别网络设计

桑海峰[1],赵子裕[1],何大阔[2]
(1. 沈阳工业大学信息科学与工程学院,辽宁沈阳 110870;2. 东北大学信息科学与工程学院,辽宁沈阳 110819)

摘 要: 视频帧中复杂的环境背景、照明条件等与行为无关的视觉信息给行为空间特征带来了大量的冗余和噪声,一定程度上影响了行为识别的准确性.针对这一点,本文提出了一种循环区域关注单元以捕捉空间特征中与行为相关的区域视觉信息,并根据视频的时序特性又提出了循环区域关注模型.其次,本文又提出了一种能够突显整段行为视频序列中较为重要帧的视频帧关注模型,以减少异类行为视频序列间相似的前后关联给识别带来的干扰.最后,提出了一个能够端到端训练的网络模型:基于循环区域关注和视频帧关注的视频行为识别网络(Recurrent Region Attention and Video Frame Attention based video action recognition Network,RFANet).在两个视频行为识别基准UCF101数据集和HMDB51数据集上的实验表明,本文提出的端到端网络RFANet能够可靠地识别出视频中行为的所属类别.受双流结构启发,本文构建了双模态RFANet网络.在相同的训练环境下,双模态RFANet网络在两个数据集上达到了最优的性能.

关键词: 行为识别;循环区域关注;视频帧关注;循环神经网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2020)06-1052-10
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.06.002

Recurrent Region Attention and Video Frame Attention Based Video Action Recognition Network Design

SANG Hai-feng[1],ZHAO Zi-yu[1],HE Da-kuo[2]
(1. *School of Information Science & Engineering,Shenyang University of Technology,Shenyang,Liaoning* 110870,*China*;
2. *College of Information Science & Engineering,Northeastern University,Shenyang,Liaoning* 110819,*China*)

Abstract: In video frames,the complex environment background,lighting conditions and other visual information unrelated to action bring a lot of redundancy and noise to action spatial feature,which affects the accuracy of action recognition to some extent. In view of this,this paper proposes a recurrent region attention cell to capture the visual information of the region related to the action in spatial features. Based on the sequence nature of video,a recurrent region attention model (RRA) is proposed. Secondly,this paper proposes a video frame attention model (VFA) that can highlight the more important frames in the video sequence of the whole action,so as to reduce the interference brought by the similar before and after correlation between video sequences of different actions. Finally,this paper presents a network model which can perform end-to-end training:recurrent region attention and video frame attention based video action recognition network (RFANet). Experiments on two video action recognition benchmark UCF101 dataset and HMDB51 dataset show that the RFANet proposed in this paper can reliably identify the category of action in the video. Inspired by the two-stream structure,we construct a two-modalities RFANet network. In the same training conditions,the two-modalities RFANet network achieved optimal performance on both datasets.

Key words: action recognition;recurrent region attention;video frame attention;recurrent neural network

1 引言

卷积神经网络能够在大规模监督数据集的帮助下学习原始视觉数据的判别性空间表示,近年来,凭借其出色的建模能力,在静态图像领域的识别与分类任务中取得了巨大的成功[1],也逐渐引入到视频领域解决基于视频的行为识别问题.

视频行为识别一直是计算机视觉领域的研究热

收稿日期:2019-06-03;修回日期:2019-11-22;责任编辑:梅志强
基金项目:国家自然科学基金(No.61773105,No.61374147);辽宁省自然科学基金(No.20170540675);辽宁省教育厅科研项目(No.LQGD2017023)

点,其目标是分析一个未知视频或者图像序列中正在进行的行为. 近些年,已经提出了许多基于双流 2D CNN 的方法来提高视频行为识别的准确性. 然而,这些方法对视频序列时间上下文的访问是有限的,主要因为空间流 2D CNN 仅应用于单帧 RGB 视频帧,时间流 2D CNN 仅应用于单堆光流堆叠,忽略了视频在时间上的时序特性.

针对上述问题,Wang 等人[2]基于长期时间结构并结合稀疏时间采样策略,提出了时序分割网络(Temporal Segment Networks,TSN). TSN 也是由空间流网络和时间流网络组成,但不同于之前的双流网络,TSN 使用从整个视频序列中稀疏采样得到的一系列短片段作为输入. 每个短片段都将通过网络获得其关于行为类别的初步预测,并通过"段共识函数"得到关于整段视频的视频级预测结果.

时间流网络的输入主要是基于原始行为视频得到的光流特征. 光流特征提供了高质量的运动信息,它的引入给行为识别的准确性带来了显著的提升. 获取运动信息的另一种方法是 Tran 等人[3]提出的 C3D 网络,通过 3D 卷积核提取视频序列的空间和时间特征,以弥补 2D CNN 在时间维度上的不足.

使用卷积神经网络识别视频中行为的方法主要聚焦在固定长度视频的短期模式上,难以直接捕获可变长度视频的长期模式. 循环神经网络特别是长短期记忆(LSTM)[4]神经网络,被认为是处理长期序列数据的有效模型. Donahue 等人[5]结合 CNN 和 LSTM 提出了一种能够端到端训练的 LRCN 模型,采用 RGB 模态和光流模态[6]作为网络的输入,最后对两个模态模型的输出取平均得到最终分类结果.

将卷积神经网络和 LSTM 网络结合起来[5],直接在数据集上进行端到端的联合训练能够更好地学习行为视频序列的时空信息. 但是,视频帧中行为所处环境复杂,行为主体比例、照明条件等变化频繁,给空间信息带来了冗余和噪声;其次,不同类别的行为视频在时序上可能存在着相似的前后关联,使得 LSTM 网络预测失误.

简单地结合卷积神经网络和 LSTM 网络的方法,忽略了与行为无关的环境背景、照明条件等视觉信息在空间上带来的冗余和噪声以及异类行为视频帧在时序上的相似前后关联给识别带来的干扰. 针对这样的问题,本文决定对行为视频序列的空间信息和时序信息加以关注.

人类观看视觉图像时,视觉系统并不是同时处理整个图像的,而是通过对全局图像进行快速扫描,获取需要重点关注的目标区域的. 而后对这一区域投入更多关注力资源,以获取所需关注目标的更多细节信息并抑制其他无用信息对当前目标的影响. 人类的这种视觉关注机制极大地提高了对视觉信息处理的效率与准确性.

受人类视觉关注机制启发,Xu 等人[7]在图像字幕任务中引入了软关注机制. 随后,该软关注机制被应用到视频分析任务中. Sharma 等人[8]提出了一种基于多层循环神经网络的软关注 LSTM 模型,该模型对视频序列中的部分视频帧进行选择性关注,以提高模型识别视频中行为的能力. Yan 等人[9]结合分层多尺度 RNN 和关注机制,提出了一种新颖的分层多尺度关注网络. 该方法使用新提出的随机神经元梯度估计方法,即 Gumbel-softmax,来实现时间边界检测器和随机硬关注机制. Yu 等人[10]通过使用联合时空关注模型提出了一种新颖的高级行为表示. 该方法建立空间卷积(2D)分支以获得空间关注引导并构建额外的时间卷积(1D)分支,将两个分支集成到时空单元中,并且通过 softmax 函数获得空间关注门. 最后,应用两级全局关注分支来获得更好的空间关注引导.

然而,这些关注模型均与循环神经网络(LSTM、GRU)高度集成,计算过程复杂,给网络的训练带来了昂贵的计算代价. 为了避免关注模型的增添给网络的训练过程带来沉重的计算负担,本文分别为行为视频序列的空间信息和时序信息提出了新颖、简单且有效的关注模型.

为了解决上述挑战,本文做出了如下贡献:(1)提出了一种循环区域关注单元,该单元可以有效地捕捉视频帧空间特征中与行为相关的区域视觉信息,以减小冗余信息和噪声信息对行为空间特征的干扰. 然后针对视频的时序特性,基于循环区域关注单元,提出了循环区域关注模型(Recurrent Region Attention model,RRA). 该模型中的循环区域关注单元按照视频的时序进行迭代,使得循环区域关注模型的关注性能逐步提升. (2)提出了一种视频帧关注模型(Video Frame Attention model,VFA)来突显整段视频序列中较为重要的帧,以降低行为视频序列的类间相似性带来的干扰. (3)构建了一个能够端到端训练的基于循环区域关注和视频帧关注的视频行为识别网络(Recurrent Region Attention and Video Frame Attention based video action recognition network,RFANet).

本文采用稀疏时间采样策略[2],在行为视频序列上获取行为视频序列子集,作为 RFANet 网络的输入,使得 RFANet 能够建模整个视频序列的长期时间模式.

受双流结构启发,本文将 RGB 模态 RFANet 和光流模态 RFANet 的视频级预测结果进行融合以产生最终的行为类别预测结果. 双模态 RFANet 网络结构如图 1 所示. 首先对行为视频序列进行时序分割,并在通过时

序分割得到的每个片段中采样多个连续帧得到关于每个片段的段输入序列. 所有的段输入序列的集合则为单模态 RFANet 网络的输入序列. 对每个模态的输入序列进行如下操作:使用卷积神经网络获取输入序列中每个段输入序列的空间特征序列,然后使用本文提出的循环区域关注模型(RRA)捕捉每段空间特征序列中与行为相关的区域视觉信息. 堆叠所有通过循环区域关注得到的段空间特征序列,并送入视频帧关注模型(VFA). 再使用双向 LSTM 在每一时刻对视频帧关注后得到的行为空间特征序列做出预测,并使用段共识函数获取视频级预测结果. 最后将两个模态的视频级预测结果进行融合,得到关于行为视频序列类别的最终预测. 实验结果表明,在两个行为识别基准数据集上,双模态 RFANet 网络达到了较高的识别精度.

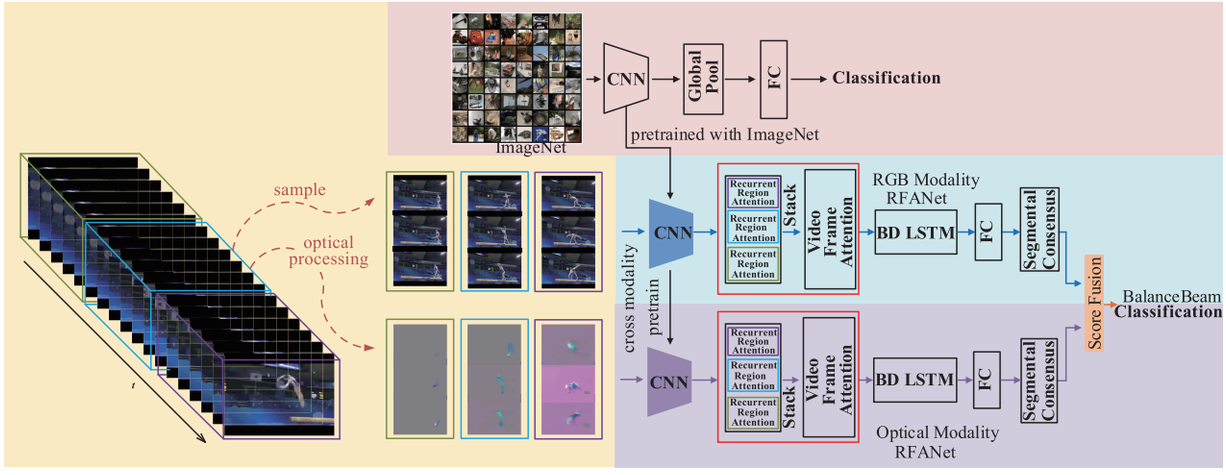


图1 双模态RFANet网络, 其中右半部分的红色框为本文提出的两级关注模型

2 基于循环区域关注和视频帧关注的视频行为识别网络的设计

在本节中详细描述了本文提出的基于循环区域关注和视频帧关注的视频行为识别网络(RFANet)模型,并介绍了本文使用的双模态融合方式,图 2 展示了 RFANet 网络的模型结构.

本文提出的 RFANet 主要由卷积神经网络、两级关注和双向 LSTM 三部分组成. 下面将依次对本文提出的 RFANet 网络模型进行详细的展开说明.

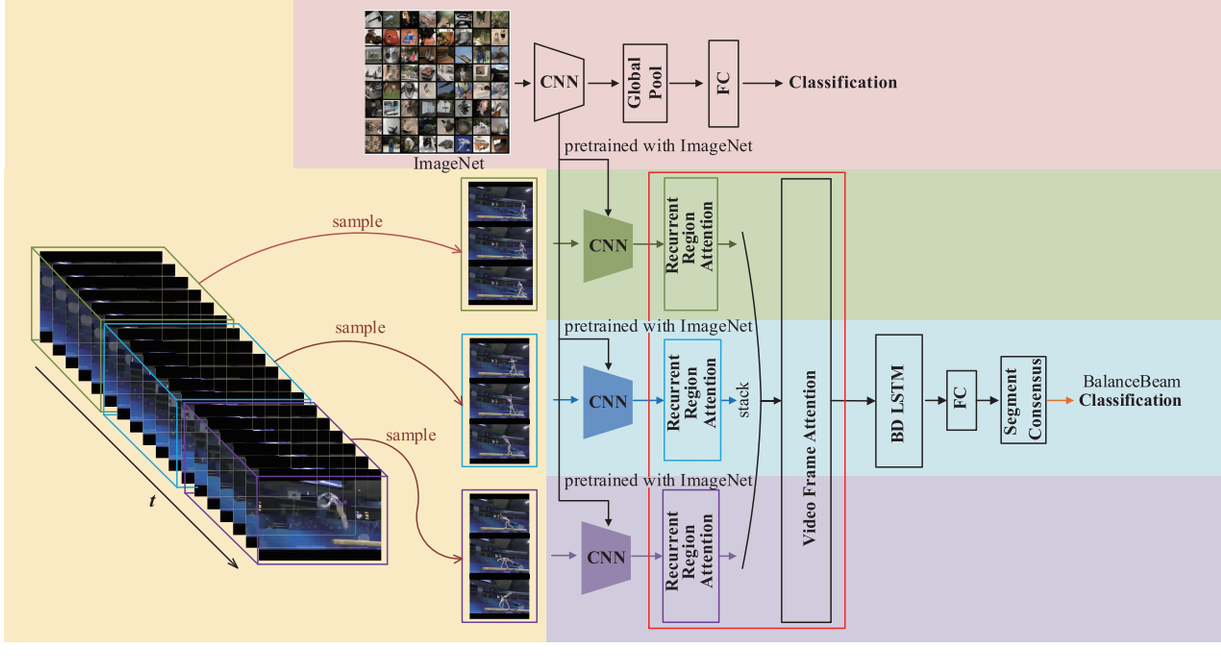


图2 基于循环区域关注和视频帧关注的视频行为识别网络模型

2.1　网络输入

假设某一视频序列 $\boldsymbol{V}$ 共有 T 帧,本文把此视频序列表示为:

$$\boldsymbol{V}=(\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_t,\cdots,\boldsymbol{F}_T),t\in T \tag{1}$$

然后使用 Wang 等人[7]提出的时序分割思想,将整个行为视频序列等间隔分割成 I 份片段,并在每份片段里随机采样 l 个连续视频帧,采集到的视频子序列 $\boldsymbol{v}$ 作为 RFANet 网络的输入,表示为:

$$\boldsymbol{v}=(\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_i,\cdots,\boldsymbol{P}_l,\boldsymbol{P}_{l+1},\boldsymbol{P}_{l+2},\cdots,\boldsymbol{P}_{l+i},\cdots,\boldsymbol{P}_{2l},\cdots,\boldsymbol{P}_{I\times l}),\ \boldsymbol{v}\in\boldsymbol{V},i\in l,I\times l\in T \tag{2}$$

2.2　循环区域关注模型

本文首先提出了一种循环区域关注单元(recurrent region attention cell),这种循环区域关注单元可以捕捉视频帧空间特征中与行为相关的区域视觉信息,进而减小了冗余信息和噪声信息对行为空间特征的影响. 循环区域关注单元的特点主要在于其内部,该神经元的输出不仅传递给下一层神经元,而且还传递回本层的循环区域关注单元. 这种传递方式构成了数据的循环. 图 3 详细的展示了循环区域关注单元的内部结构及数据流动.

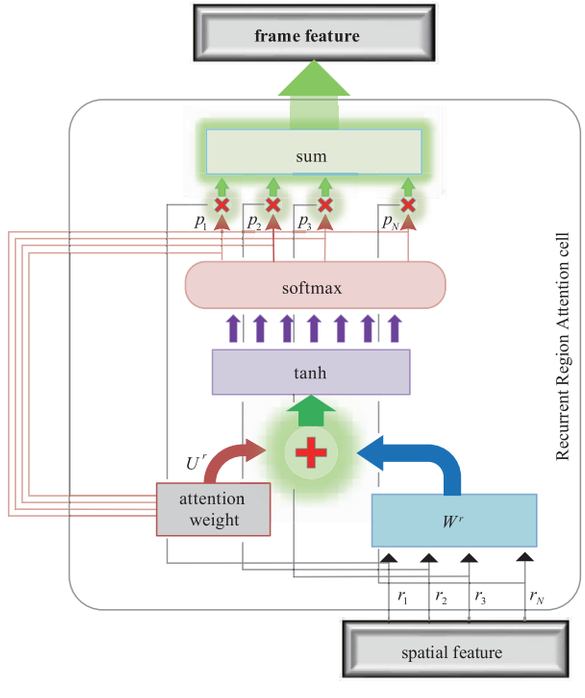


图3　循环区域关注单元内部结构图

受 LSTM 网络中 LSTM 单元的迭代工作方式启发,根据视频的时间序列性质,在循环区域关注单元的基础上,本文提出了一种循环区域关注模型(RRA),该模型具有时间顺序性质. 因此,可以通过视频帧之间的时间顺序更好地捕获与行为相关的区域视觉信息. RRA 中的循环区域关注单元根据视频的时序进行迭代,因此 RRA 对与行为相关的区域视觉信息的关注性能逐渐提高. 不同于机器翻译、语音识别、图像标注等其他领域中使用的结合文本特征和循环神经网络隐藏状态的软关注机制[11],本文提出的循环区域关注模型是一种以图像空间特征和区域关注权重作为输入,仅依靠循环区域关注单元在层内神经元之间形成有向环连接的神经网络模型. 为了方便表达循环区域关注模型的"循环"特性,图 4 将其按照输入序列的长度展开. 展开后,循环区域关注模型可以看作是由一系列核心模块(循环区域关注单元)组成的阵列. 在该阵列中,前一帧(前一时刻)与后一帧(后一时刻)的循环区域关注单元相互连接.

使用卷积神经网络提取视频子序列 $\boldsymbol{v}$ 中某一视频帧 $\boldsymbol{P}_i$ 的空间特征 $\boldsymbol{f}_i$,其形状为:[height,width,channel]. 其中 height 表示空间特征 $\boldsymbol{f}_i$ 的高,width 表示空间特征 $\boldsymbol{f}_i$ 的宽,channel 表示空间特征 $\boldsymbol{f}_i$ 的通道数.

对于视频子序列 $\boldsymbol{v}$ 中的某一帧 $\boldsymbol{P}_i$,我们能够得到关于这一帧的空间区域特征序列 $(\boldsymbol{r}_1^i,\boldsymbol{r}_2^i,\cdots,\boldsymbol{r}_N^i)$,其中 N 表示视频帧 $\boldsymbol{P}_i$ 的空间特征区域总数,

$$N=\text{height}\times\text{width} \tag{3}$$

将这些区域特征赋予权重求和便得到循环区域关注后的视频帧空间特征:

$$\boldsymbol{f}_i^r=\sum_{j=1}^{N}\boldsymbol{p}_j^i\boldsymbol{r}_j^i \tag{4}$$

则通过循环区域关注后,视频子序列 $\boldsymbol{v}$ 的空间特征序列表示为 $\boldsymbol{v}^r=(\boldsymbol{f}_1^r,\boldsymbol{f}_2^r,\cdots,\boldsymbol{f}_i^r,\cdots,\boldsymbol{f}_{I\times l}^r)$.

其中,$\boldsymbol{r}_j^i$ 表示第 i 帧的第 j 个区域特征,$\boldsymbol{p}_j^i$ 是与 $\boldsymbol{r}_j^i$ 相对应的第 i 帧中第 j 个区域的关注权重,$\boldsymbol{f}_i^r$ 表示的是通过循环区域关注捕捉了与行为相关的区域视觉信息后的视频帧空间特征.

$\boldsymbol{p}_j^i$ 由下列等式(5)、(6)计算得到:

$$\boldsymbol{u}_j^i=\boldsymbol{\omega}^r\tanh(\boldsymbol{W}^r\boldsymbol{r}_j^i+\boldsymbol{U}^r\boldsymbol{p}_j^{i-1})+b^r \tag{5}$$

$$\boldsymbol{p}_j^i=\frac{\exp(\boldsymbol{u}_j^i)}{\sum_{j=1}^{N}\exp(\boldsymbol{u}_j^i)} \tag{6}$$

其中,$\boldsymbol{W}^r$、$\boldsymbol{U}^r$、$\boldsymbol{\omega}^r$、b^r 为循环区域关注模型学习到的共享参数. 第 i 帧中与行为有关的区域视觉信息则由 $\boldsymbol{p}_j^{i-1}$ 进行捕捉,这样可以减少行为所处环境中复杂背景、照明条件等与行为无关的视觉信息对识别带来的干扰. $\boldsymbol{p}_j^{i-1}$ 表示当前被关注帧的前一帧的关注权重. 因此,第 i 帧的区域关注权重是由当前帧空间特征和前一帧区域关注权重共同决定的. 本文希望能够通过这样的循环区域关注,减少行为所处环境中噪声和冗余信息给识别准确性带来的影响,同时解决帧与帧之间的空间相似

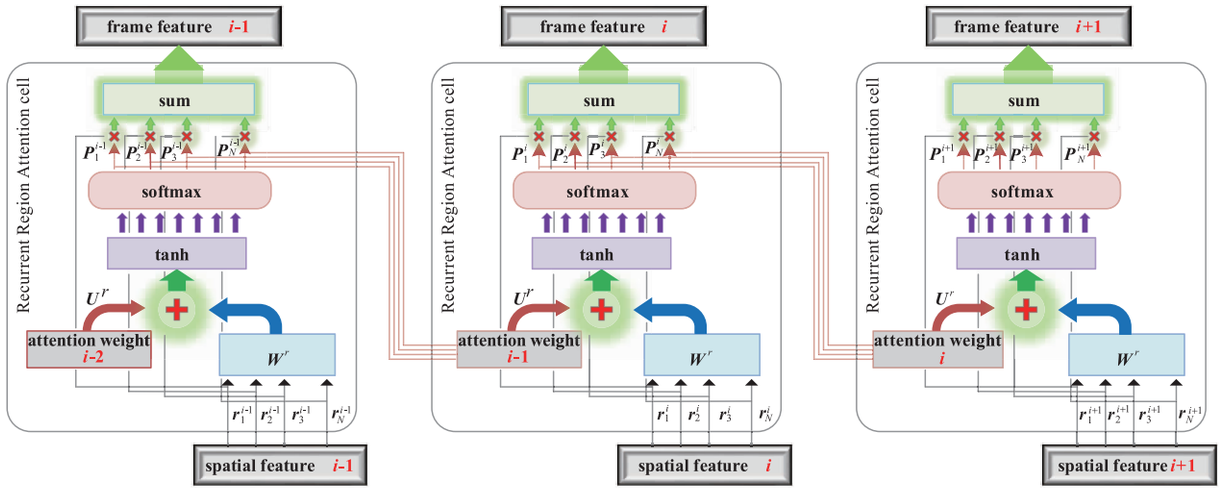


图4 循环区域关注模型按输入序列长度展开图。当前帧(当前时刻)的空间特征与前一帧(前一时刻)的区域关注权重作为当前时刻循环区域关注单元的输入。经过循环区域关注单元内部的softmax函数得到的概率即为当前帧区域特征的关注权重。利用该权重对区域特征加权求和便获得能够突出行为视觉信息的视频帧空间特征

性带来的特征模糊问题,使得空间特征更具有正确表征行为的能力.

2.3 双向循环神经网络

本文的网络结构是在循环神经网络的 LSTM 单元[4]基础上建立的. 图 5 展示了 LSTM 单元的内部结构.

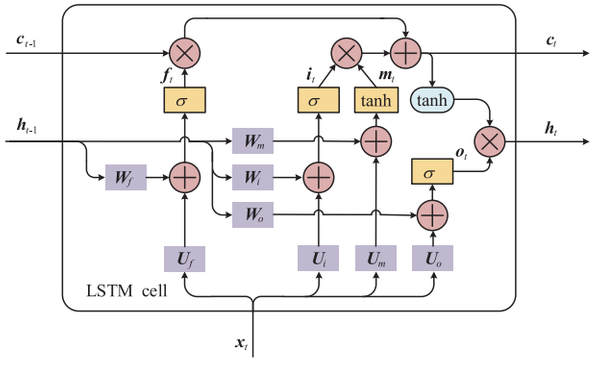


图5 LSTM单元内部结构图

给定 LSTM 单元一组特征序列($\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_{I\times l}$)作为输入,重复式(7)~(12),便可得到与之对应的隐藏状态($\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_{I\times l}$).

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{\mathrm{f}}\boldsymbol{h}_{t-1}+\boldsymbol{U}_{\mathrm{f}}\boldsymbol{x}_t+b_{\mathrm{f}}) \tag{7}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{\mathrm{i}}\boldsymbol{h}_{t-1}+\boldsymbol{U}_{\mathrm{i}}\boldsymbol{x}_t+b_{\mathrm{i}}) \tag{8}$$

$$\boldsymbol{m}_t=\tanh(\boldsymbol{W}_{\mathrm{m}}\boldsymbol{h}_{t-1}+\boldsymbol{U}_{\mathrm{m}}\boldsymbol{x}_t+b_{\mathrm{m}}) \tag{9}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\boldsymbol{m}_t \tag{10}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{\mathrm{o}}\boldsymbol{h}_{t-1}+\boldsymbol{U}_{\mathrm{o}}\boldsymbol{x}_t+b_{\mathrm{o}}) \tag{11}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{12}$$

其中,$\boldsymbol{W}$、$\boldsymbol{U}$、b 为网络学习到的共享参数. $\sigma(\cdot)$表示 sigmoid 函数,$\odot$表示按元素相乘.

视频序列中的行为具有前后连贯性,因此仅根据过去时刻的行为信息对当前时刻做出预测并不严谨. LSTM 单元仅考虑了过去时刻的输入对当前时刻输出的影响,然而本文需要的输出是依赖于整个视频序列的. 为了考虑未来时刻行为信息对当前时刻的影响,本文采用双向 RNN[12] 的思想,结合正向 LSTM 和反向 LSTM 组成双向 LSTM,充分利用视频序列中的前后依赖性. 图 6 为双向 LSTM 按时刻展开图.

当前时刻输入 $\boldsymbol{x}_t$ 分别输入正向 LSTM 和反向 LSTM,对应产生当前时刻的正向隐藏状态 $\boldsymbol{h}_t^{\mathrm{F}}$ 和反向隐藏状态 $\boldsymbol{h}_t^{\mathrm{B}}$,则此时双向 LSTM 的输出表示为 $\boldsymbol{h}_t=[\boldsymbol{h}_t^{\mathrm{F}},\boldsymbol{h}_t^{\mathrm{B}}]$,$[\cdot]$表示拼接两个隐藏状态.

2.4 视频帧关注模型

虽然双向 LSTM 能够充分学习某类行为视频序列的前后关联性,但是异类行为视频序列之间可能在时序的某些部分存在多数连续的相似帧,这便导致双向 LSTM 因异类行为视频的空间特征序列在时序上相似的前后关联预测失误. 因此,本文又提出了一种能够突显整段视频序列中较为重要帧的视频帧关注(Video Frame Attention,VFA),以减少行为视频序列的类间相似性带来的干扰. 本文提出的 VFA 采用 RRA 的堆叠输出作为输入并与 LSTM 单元分离,不仅减少了计算量,还不降低网络的计算速度,同时又提高了网络的性能. RRA 和 VFA 之间的数据传输过程如图 2 所示. 视频帧关注结构如图 7 所示.

给定该关注关于视频子序列 $\boldsymbol{v}$ 的空间特征序列 $\boldsymbol{v}^r=(\boldsymbol{f}^r_1,\boldsymbol{f}^r_2,\cdots,\boldsymbol{f}^r_i,\cdots,\boldsymbol{f}^r_{I\times l})$,经过下述等式(13),便获得视频帧关注后的空间特征序列 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_{I\times l})$.

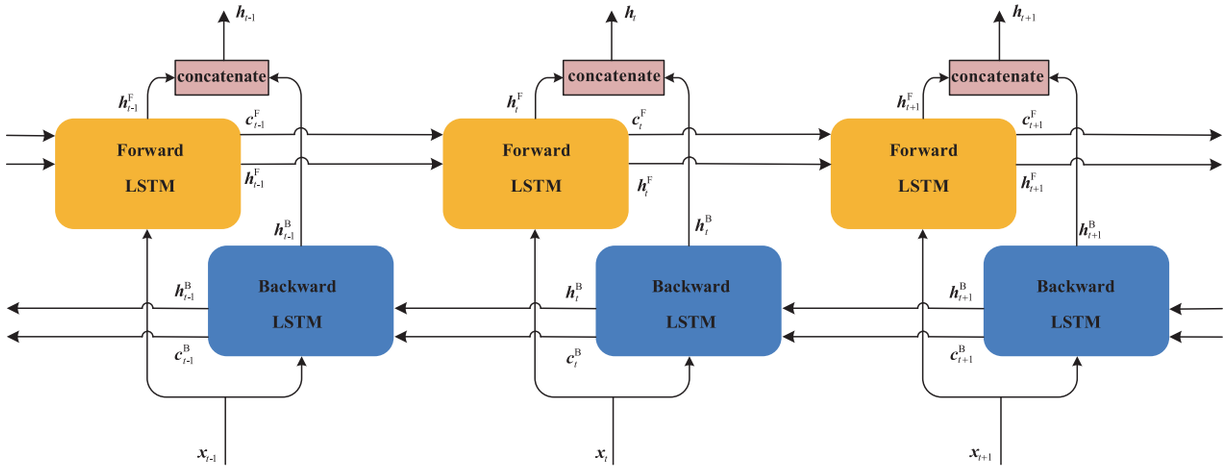


图6　双向LSTM按时刻展开图

$$\boldsymbol{x}=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}^{r} \tag{13}$$

其中,$\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_i,\cdots,\boldsymbol{\alpha}_{l\times l})$是视频子序列 $\boldsymbol{v}$ 的空间特征序列 $\boldsymbol{v}^r$ 的帧权重,由式(14)、(15)得到:

$$\boldsymbol{q}=\boldsymbol{u}^{\mathrm{T}}\tanh(\boldsymbol{W}^{v}\boldsymbol{v}^{r}+b^{v}) \tag{14}$$

$$\boldsymbol{\alpha}=\mathrm{softmax}(\boldsymbol{q}) \tag{15}$$

其中,$\boldsymbol{W}^v$、b^v、$\boldsymbol{u}$ 为网络学习到的共享参数. 本文希望能够通过视频帧关注,减少异类行为视频序列在时序上的相似前后关联给双向 LSTM 学习序列前后依赖关系带来的干扰.

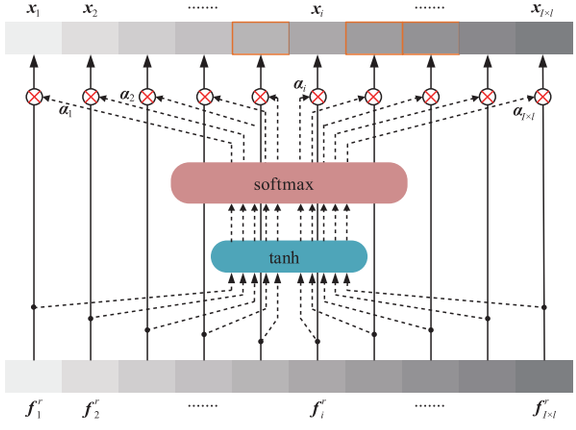


图7　视频帧关注结构

2.5　视频级预测

不同于以往将 LSTM 最后时刻的输出用于分类,本文首先将双向 LSTM 各时刻的输出均传递给全连接层,在每一时刻做出关于行为类别的初步预测,然后延续 Wang 等人[7]提出的段共识函数等式(16),使得序列中各时刻的初步预测结果达成共识,生成视频级预测结果:

$$\hat{\boldsymbol{Y}}=\mathrm{Consensus}(\hat{\boldsymbol{y}}_1,\hat{\boldsymbol{y}}_2,\cdots,\hat{\boldsymbol{y}}_{l\times l}) \tag{16}$$

2.6　参数更新

假设训练集中有 K 个视频序列样本,本文将第 i 个视频序列样本的真实类别标签表示为 $\boldsymbol{Y}_i$,其视频级预测类别表示为 $\hat{\boldsymbol{Y}}_i$,我们可以通过最小化代价函数式(17)对网络参数进行更新:

$$J=\frac{1}{K}\sum_{i=1}^{K}\left[-\boldsymbol{Y}_i^{\mathrm{T}}\log(\hat{\boldsymbol{Y}}_i)-(1-\boldsymbol{Y}_i)^{\mathrm{T}}\log(1-\hat{\boldsymbol{Y}}_i)\right] \tag{17}$$

2.7　模态融合

为了充分利用行为视频序列提供的时空信息和运动信息,本文考虑融合 RGB 模态 RFANet 网络和光流模态 RFANet 网络的视频级预测结果. 由于本文提出的 RFANet 是端到端的网络,所以本次工作中选用概率融合的方式对不同模态 RFANet 网络的视频级预测结果进行融合.

本文首先为每个模态的输入序列训练一个单独的模型,然后对属于不同模态的独立模型的视频级预测结果进行概率融合. 本文将 RGB 模态 RFANet 网络的视频级预测结果表示为$\hat{\boldsymbol{Y}}_{\mathrm{RGB}}$,将光流模态 RFANet 网络的视频级预测结果表示为$\hat{\boldsymbol{Y}}_{\mathrm{optical}}$,则通过概率融合方法得到的最终预测结果可由式(18)得到. 其中 λ 表示 RGB 模态 RFANet 网络视频级预测结果$\hat{\boldsymbol{Y}}_{\mathrm{RGB}}$的融合权重. 具体的概率融合将在第 3 节的实验部分中进行详细说明.

$$\hat{\boldsymbol{Y}}_{\mathrm{Fusion}}=\lambda*\hat{\boldsymbol{Y}}_{\mathrm{RGB}}+(1-\lambda)*\hat{\boldsymbol{Y}}_{\mathrm{optical}} \tag{18}$$

3　实验

在本节中,首先介绍评估数据集和本文提出的方法的实现细节,然后验证本文提出的两级关注在提高识别精度上的有效性以及 RFANet 网络的识别性能,最后对双模态 RFANet 网络的性能进行了评估并与 SOTA 方法进行了比较.

3.1　数据集

本文在两个流行的视频行为识别数据集上评估并比较所提出的 RFANet 网络的性能.

UCF101 数据集[13]是从 YouTube 上收集到的具有 101 个类别的视频行为数据集,共有 13320 个视频,平均每个视频长度为 180 帧. UCF101 在行为类别上提供了最大的多样性,涉及范围包括生活中的日常活动乃至极限运动,并且视频中存在相机运动,物体比例、背景、照明条件等变化也较大,是具有挑战性的数据集.

HMDB51 数据集[14]是由来自各种来源(如电影和 YouTube 视频)的 6766 个视频片段组成的视频行为数据集,具有 51 个行为类别.

本文将遵循两个数据集的原始评估方案,报告三个训练/测试 split 的平均准确度.

3.2 实验细节

在配有 2 块 NVIDIA RTX2080Ti GPU 和 32GB 内存的计算机上使用 Tensorflow 框架完成本次工作.

以 UCF101 数据集的实验为例. 实验中,采用稀疏采样策略[2]对每个模态的视频序列进行采样,将采样得到的视频子序列作为单模态 RFANet 网络的输入. 本文使用 BN Inception[1]网络提取视频帧的空间特征,使用 Momentum 优化算法优化网络参数. 其中 Batchsize 大小设置为 32,动量(momentum)设置为 0.9,网络输入维度设置为 15,段序列长度设置为 5,即将视频序列等间隔分割成 3 段,每个片段内采样 5 个连续视频帧. 使用均值段共识函数获取 RGB 模态 RFANet 网络或光流模态 RFANet 网络的最终视频级预测结果. 其中卷积神经网络的输入维数是(15,224,224,3),输出维数为(15,7,7,1024),这里的 15 表示由稀疏采样策略采样得到的 15 帧,其中每一帧的空间特征维数为(7,7,1024),分别对应单帧空间特征的 height、width、channel,则空间区域特征的个数为 $N=7\times7=49$,即每一个视频帧经过卷积之后得到的空间特征有 49 个区域,每一个区域特征的维数为(1,1,1024).

对于 RGB 模态的 RFANet 网络,本文使用来自 ImageNet[15]的预训练模型初始化卷积网络权值. 视频帧关注中的隐藏单元数设置为 256,LSTM 隐藏单元的数量设置为 256,整个训练过程在 120 个 epoch 中停止. 对于光流模态的 RFANet 网络,由于光流图分布与 RGB 图像不同,本文利用线性变换将光流场离散为与 RGB 图像范围相同的 0 ~ 255 区间,然后使用交叉模态预训练的方法利用 RGB 模态网络中的卷积网络初始化光流模态网络中卷积神经网络的权值,减少了光流模态网络的训练时长并适当的避免过拟合. 视频帧关注中的隐藏单元数设置为 256,LSTM 隐藏单元的数量设置为 256,最大迭代设置为 250 个 epoch. 为了避免在训练过程中因模型复杂导致过拟合,本文采用随机裁剪、水平翻转、角裁剪和尺度抖动等技术进行数据增强,并在双向 LSTM 后增加了一个 Dropout 层. RGB 模态 RFANet 网络的 dropout rate 设置为 0.5,光流模态 RFANet 网络的 dropout rate 设置为 0.7.

为了加快训练,本文使用了多 GPU 并行策略. UCF101 数据集在 RGB 模态 RFANet 网络的训练时间为 50 小时左右,在光流模态 RFANet 网络的训练时间为 90 小时左右.

3.3 性能评估

3.3.1 RFANet 网络性能评估

在本小节中,首先验证了本文提出的循环区域关注和视频帧关注在提高网络识别性能上的有效性,然后验证了单模态 RFANet 网络的识别性能.

本文将 BN Inception[1]在 ImageNet[15]数据集上学习到的模型权值迁移到了 UCF101 数据集上. 为了公平比较,本文使用预训练的方法仅在 UCF101 数据集的 RGB 模态输入上进行实验,以突显本文提出的循环区域关注和视频帧关注给识别性能带来的提升. 实验结果总结在表 1 中. 识别精度为各种方法在 UCF101 数据集的三个训练/测试 split 上的平均准确度,网络模型根据预训练数据集进行分组. 表中最后三行分别为本文的基础模型 ConvNet + BDLSTM、加入了本文提出的循环区域关注的网络模型 ConvNet + RRA + BDLSTM 和加入了本文提出的两级关注的最终网络模型 RFANet. 循环区域关注的加入使得网络识别精度较基础模型提升 2.4%,两级关注的加入使得识别精度提升 3.5%. 通过比较三个模型的识别精度可以发现,本文提出的循环区域关注和视频帧关注给网络识别性能带来了显著的提升.

表 1 多种方法在 UCF101 数据集上的平均识别精度

Method(RGB)	Pretrain	Resolution	Backbone	Acc
Spatial Stream ResNet[24]	ImageNet	224 × 224	ResNet-50	82.3%
RGB-I3D[23]	ImageNet	224 × 224	Inception V1	84.5%
TSN[2]	ImageNet	224 × 224	BN Inception	85.7%
Multimodal Fusion Network[16]	ImageNet	224 × 224	ResNet-152	86.2%
ConvNet + BDLSTM	ImageNet	224 × 224	BN Inception	86.1%
ConvNet + RRA + BDLSTM	ImageNet	224 × 224	BN Inception	88.5%
RFANet	ImageNet	224 × 224	BN Inception	89.6%

为了更直观地反映本文所提出的循环区域关注对网络识别性能的改进,本文对循环区域关注的关注效果进行了可视化. 可视化热力图如图 8 所示. 可以清楚地观察到,本文提出的循环区域关注的关注资源主要集中在与行为相关的空间区域,从而减少了与行为无关的视觉信息带来的干扰,提升了网络的识别性能.

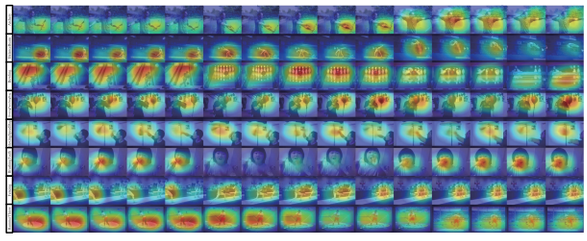


图8　循环区域关注模型关注效果的可视化热力图

然后本文又比较了表 1 中在 ImageNet 数据集上预训练的 TSN(RGB)-BN Inception[2]、Multimodal Fusion Network(RGB)-ResNet152[16] 和 RFANet(RGB)-BN Inception 的性能,并看到本文的 RFANet(RGB)在 UCF101 数据集上的表现优于 TSN(RGB) 3.9%,优于 Multimodal Fusion Network(RGB) 3.4%. 这种卓越的性能表明,在迁移学习方面仅使用 RGB 模态输入时,本文提出的 RFANet 学习到的时空表示比 TSN 和 Multimodal Fusion Network 更有效. RFANet 网络在 ImageNet 数据集的预训练设置下获得最佳性能.

3.3.2　双模态 RFANet 网络性能评估

在本小节中,本文在 UCF101 数据集和 HMDB51 数据集上验证了 RFANet 网络的双模态融合性能. 本文选用概率融合的方式对 RGB 模态 RFANet 网络和光流模态 RFANet 网络的视频级预测结果进行融合. 本文首先对概率融合的融合权重进行了研究,UCF101 数据集和 HMDB51 数据集的 3 个 split 对于不同融合权重的双模态融合结果三维折线图如图 9、10 所示. 图中 λ 表示 RGB 模态 RFANet 网络的视频级预测结果 $\hat{\boldsymbol{Y}}_{\mathrm{RGB}}$ 的融合权重,当 λ 为 0.3 时,双模态 RFANet 网络在 UCF101 数据集和 HMDB51 数据集的 3 个 split 上均达到了最优的识别性能. 因此,本文最终选择了 $\lambda = 0.3$ 的融合权重对两个模态的视频级预测结果进行融合.

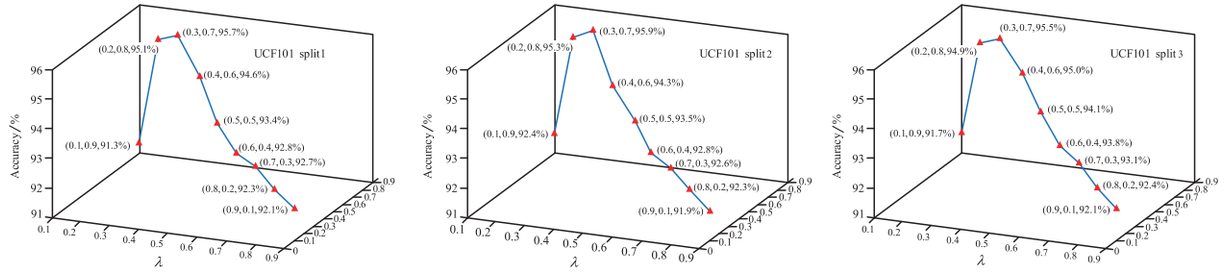


图9　UCF101数据集关于不同融合权重的双模态融合结果三维折线图

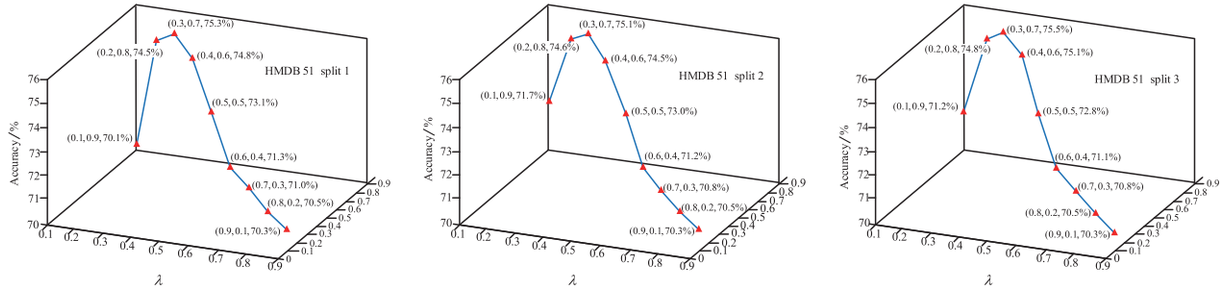


图10　HMDB51数据集关于不同融合权重的双模态融合结果三维折线图

最后,本文分别在 UCF101 数据集和 HMDB51 数据集上对本文提出的 RFANet 网络的双模态融合性能与当前 SOTA 方法的性能进行了比较. 结果如表 2 所示. 在 UCF101 数据集上,本文提出的 RFANet 网络的表现优于 TSN[2] 1.7%,优于 Multimodal Fusion Network[16] 和 ResNet + TSN[17] 0.9%,优于 TVNets + IDT[18] 0.3%,超过目前 SOTA 方法的性能;在 HMDB51 数据集上的表现优于 TSN[2] 6.8%,优于 ResNet + TSN[17] 3.5%,优于 TVNets + IDT[18] 2.7%,优于 Pillar Networks ++[19] 1.7%. 这表明本文提出的 RFANet 具有良好的泛化能力.

表 2 RFANet 与 SOTA 方法的双模态性能比较

Method	UCF101	HMDB51
DT + MVSV[25]	83.5%	55.9%
iDT + HSV[20]	87.9%	61.1%
VideoLSTM[21]	89.2%	–
C3D[3]	85.2%	51.6%
Two Stream + LSTM[10]	88.6%	–
ConvNet + LSTM[5]	82.3%	–
TDD + FV[22]	90.3%	63.2%
HAN[9]	92.7%	64.3%
JSTA[10]	93.7%	65.3%
TSN (2 modalities)[2]	94.0%	68.5%
Multimodal Fusion Network[16]	94.8%	–
ResNet + TSN[17]	94.8%	71.8%
TVNets + IDT[18]	95.4%	72.6%
Pillar Networks + +[19]	–	73.6%
RFANet (2 modalities)	95.7%	75.3%

4 结论

本文提出的循环区域关注和视频帧关注给网络识别性能带来了显著的提升. 虽然双模态 RFANet 网络的识别性能在两个基准数据集上达到了新的技术水平,但是由于本文的 RFANet 网络是端到端的网络,在双模态融合时本文只使用了概率融合的融合方式. 在这种融合方法中每个模态的模型只能访问当前模态的特征,无法学习不同模态间的交互. 在后续的工作中,还将尝试在模型的不同位置处使用不同的方式进行更全面的模态融合,以达到更高的识别性能.

参考文献

[1] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[A]. International Conference on Machine Learning[C]. Lille, France: International Machine Learning Society, 2015. 448 – 456.

[2] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition [A]. European Conference on Computer Vision[C]. Amsterdam, Netherlands: Springer International Publishing, 2016. 20 – 36.

[3] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[A]. International Conference on Computer Vision[C]. Santiago, Chile: IEEE, 2015. 4489 – 4497.

[4] Hochreiter S, Schmidhuber J. Longshort-term memory[J]. Neural Computation, 1997, 9(8): 1735 – 1780.

[5] Donahue J, Hendricks L A, Guadarrama S, et al. Long-termrecurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 677 – 691.

[6] Brox T, Bruhn A, Papenberg N, et al. High accuracy optical flow estimation based on a theory for warping[J]. Computer Vision, 2004, 3024(10): 25 – 36.

[7] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention[A]. International Conference on Machine Learning[C]. Lille, France: International Machine Learning Society, 2015. 2048 – 2057.

[8] Sharma S, Kiros R, Salakhutdinov R. Action Recognition Using Visual Attention[DB/OL]. https://arxiv.org/abs/1511.04119, 2015-11-12.

[9] Yan S, Smith J S, Lu W, et al. Hierarchical multi-scale attention networks for action recognition[J]. Signal Processing: Image Communication, 2018, 61: 73 – 84.

[10] T Yu, C Guo, L Wang, et al. Joint spatial-temporal attention for action recognition[J]. Computer Science, 2018, 112(2018): 226 – 233.

[11] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[DB/OL]. https://arxiv.org/abs/1409.0473, 2014-09-01.

[12] Schuster M, Paliwal KK. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(17): 2673 – 2681.

[13] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[[DB/OL]]. https://arxiv.org/abs/1212.0402, 2012-12-03.

[14] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[A]. International Conference on Computer Vision[C]. Barcelona, Spain: IEEE, 2011. 2556 – 2563.

[15] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database[A]. Computer Vision and Pattern Recognition[C]. Miami, FL, USA: IEEE, 2009. 248 – 255.

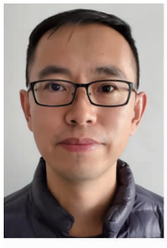
[16] Xiang L, Chuang G, et al. Multimodal keyless attention fusion for video classification[A]. 32nd AAAI Conference on Artificial Intelligence[C]. New Orleans, Louisiana, USA: AAAI, 2018. 7202 – 7209.

[17] Yuan Y, Wang D, Wang Q. Memory-Augmented Temporal Dynamic Learning for Action Recognition[DB/OL]. https://arxiv.org/abs/1904.13080, 2019-4-30.

[18] Fan L, Huang W, Gan C, et al. End-to-End Learning of Motion Representation for Video Understanding[A]. Computer Vision and Pattern Recognition[C]. Salt Lake

City,UT,USA:IEEE,2018. 6016 – 6025.
[19] Sengupta B, Qian Y. Pillar Networks + +: Distributed Non-parametric Deep and Wide Networks[DB/OL]. https://arxiv. org/abs/1708. 06250,2017-08-18.
[20] Peng X,Wang L,Wang X,et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice[J]. Computer Vision and Image Understanding,2016,150(2016):109 – 125.
[21] Li Z, Gavrilyuk K, Gavves E, et al. VideoLSTM convolves,attends and flows for action recognition[J]. Computer Vision and Image Understanding, 2018, 166 (2018):41 – 50.
[22] Wang L,Yu Q,Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[A]. Computer Vision and Pattern Recognition[C]. Boston, America: IEEE,2015. 4305 – 4314.
[23] Carreira J,Zisserman A. Quo Vadis,Action Recognition? A New Model and the Kinetics Dataset[A]. Computer Vision and Pattern Recognition[C]. Hawaii, America: IEEE,2017. 6299 – 6308.
[24] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition[A]. Advances in Neural Information Processing Systems[C]. Barcelona:NIPS,2016. 3468 – 3476.
[25] Cai Z,Wang L,Peng X,et al. Multi-view super vector for action recognition[A]. Computer Vision and Pattern Recognition[C]. Columbus,America:IEEE,2014. 596 – 603.

作者简介



桑海峰　男,1978 年生于辽宁沈阳,博士,沈阳工业大学教授,主要研究方向为视觉检测技术与图像处理.
E-mail:sanghaif@163. com

赵子裕(通信作者)　男,1995 年生于辽宁沈阳,沈阳工业大学硕士研究生,主要研究方向为视觉检测技术与图像处理.
E-mail:Maikuraky1022@ outlook. com