文章编号：1003-0530(2020)08-1272-08

引入高阶注意力机制的人体行为识别

王增强 张文强 张 良
（中国民航大学天津市智能信号与图像处理重点实验室，天津 300300）

摘 要：现有的视频行为识别方法在特征提取过程中，存在忽略各个特征之间相互作用关系的问题，对近似动作的区分效果不理想。因此，提出引入高阶注意力机制的人体行为识别方法。在深度卷积神经网络中引入高阶注意力模块，通过注意力机制建模和利用复杂和高阶的统计信息，对训练过程中特征图各个部分的权重进行重新分配，从而关注局部细粒度信息，产生有区别性的关注建议，捕获行为之间的细微差异。在 UCF101 和 HMDB51 这两个人体行为数据集上的实验结果表明，与现有方法相比，识别率得到了一定的提升，验证了所提出方法的有效性和鲁棒性，提高了对近似行为的辨别能力。

关键词：深度卷积神经网络；行为识别；深度学习；高阶注意力机制
中图分类号：TP391 **文献标识码**：A **DOI**：10.16798/j.issn.1003-0530.2020.08.010

引用格式：王增强，张文强，张良. 引入高阶注意力机制的人体行为识别[J]. 信号处理，2020，36(8)：1272-1279. DOI：10.16798/j.issn.1003-0530.2020.08.010.
Reference format：Wang Zengqiang，Zhang Wenqiang，Zhang Liang. Human Behavior Recognition with High-Order Attention Mechanism[J]. Journal of Signal Processing，2020，36(8)：1272-1279. DOI：10.16798/j.issn.1003-0530.2020.08.010.

Human Behavior Recognition with High-Order Attention Mechanism

Wang Zengqiang Zhang Wenqiang Zhang Liang
(Tianjin Key Laboratory of Advanced Signal and Image Processing, Civil Aviation University of China, Tianjin 300300, China)

Abstract: Because the existing video behavior recognition methods had the problem that ignore the interactional relationship among features in the process of feature extraction, the effect of distinguishing approximate actions was poor. Therefore, a human behavior recognition method with high-order attention mechanism was proposed. A high-order attention module was introduced to a deep convolutional neural network, which modeled and utilized the complex and high-order statistics information in attention mechanism. The goal of attention was to reallocate the weight of each part of the feature map in the process of training, so as to focus on the local fine-grained information, produce the discriminative attention proposals, and capture the subtle differences among behaviors. Extensive experiments had been conducted to validate the superiority of our method for action recognition on two human behavior datasets, including UCF101 and HMDB51. The results showed that the recognition rate has been improved to a certain extent compare with the existing methods, which validated the effectiveness and robustness of the proposed method. The method effectively improves the ability to distinguish approximate behaviors.
Key words: deep convolutional neural network; behavior recognition; deep learning; high-order attention mechanism

1 引言

随着网络多媒体的迅猛发展，涌现出的视频数据体量巨大，对视频中的人体行为进行识别成为计算机视觉领域的研究热点，广泛应用于智能交通安防，自动视频分类，高级人机交互等领域[1]。对视频中的人体动作进行识别，主要有基于传统手动提取特征的方法和基于深度学习的方法。

收稿日期：2020-03-23；修回日期：2020-06-11
基金项目：国家自然科学基金(61179045)

Bobick 等人[2]采用运动能量图像[3](Motion Energy Images, MEI)和运动历史图像[4](Motion History Images, MHI)来表示视频中人的运动。Alper Yilmaz 等人[5]提出了根据时空卷[4](Space-Time Volume, STV)的方向、速度等不同属性的变化来表征动作。上述两种方法为基于整体特征[6]提取的方式,需要将运动的人体从背景中分割出来,当背景较为复杂时,特征提取效果不好。一些研究者将目光转向局部特征的提取,A. Klaser 等人[7]提出密集轨迹采样的方式,并结合在每个点上提取的方向梯度直方图[8]、光流直方图[9]等特征来进一步提升识别性能。Heng Wang 等人[10]提出了改进的稠密轨迹特征(Improving Dense Trajectories, IDT),利用视频帧之间的光流信息和 SURF(Speeded Up Robust Features)关键点进行匹配,可以较好的克服相机视角变化带来的影响,获得了更加突出的表现。手动提取特征的方法依赖于人的经验,在背景、光线、遮挡等的影响下,提取鲁棒性更强,适应性更好的特征较为不易。

随着深度卷积神经网络在图像识别任务上取得的卓越效果,人们开始将其应用于视频中的人体行为识别。Simonyan 等人[11]提出用于行为识别的双流卷积神经网络[12-15],分为空间流和时间流,使用 RGB 图像提取空间信息,使用光流来提取时间信息,将两路提取的空间和时间特征做融合,最后将上述特征送入多分类的支持向量机(Support Vector Machine, SVM)训练以进行行为类别的预测。Wang 等人[16]提出时序分割网络(Temporal Segment Networks, TSN)进一步研究了时空流融合的方法,建模长范围的视频信息,并采用视频级监督的方式,取得了较好的效果。Du Tran 等人[17]提出 3D 卷积神经网络[18-21],将原来的卷积层和池化层由二维扩展为三维,克服了 2D 卷积不能提取时间信息的缺点,可以直接对视频进行处理,处理速度得到提升,但网络参数量大,训练收敛较为不易。

近年来,受人类视觉注意力的启发,将注意力机制[22]引入神经网络引起了大家的关注。Jaderberg M 等人[23]提出空间变换网络,即空间注意力,为一种视觉注意力形式,将注意力引导到空间中的某个位置,使卷积神经网络关注视觉场中特定区域的信息。Hu J 等人[24]提出挤压和激励网络,即通道注意力,不同通道含有的信息量及重要程度是不一样的,对与关键信息相关度高的通道赋予较大的权重,其他通道赋予较小的权重。

由以上分析可知,目前的研究工作集中于特征提取与表达的不断优化,从而提高动作识别的准确率。文献[25-27]的研究表明,身体不同部分之间的相互作用关系,有利于行为类别的判定。而上述注意力机制的引入,仅仅关注于图像或特征图的某个区域或通道,不能建模图像或特征图的各个部分之间相互作用对最终识别结果产生的影响,对类间距较小的近似动作[28]区分效果较差,限制了行为识别方法准确率的提高。

针对现有动作识别方法存在的问题,通过对文献[29]的研究与实践,受 Chen B 等人在解决行人重识别任务中类似问题的启发,提出引入高阶注意力机制的卷积神经网络(Convolutional Neural Network with High-Order Attention Mechanism, HAM-CNN)。具体实现过程为,将输入视频做稀疏采样,将采样得到的 RGB 和光流图片分别送入 HAM-CNN,网络中生成的特征图在高阶注意力机制的作用下,对图上各个部分之间复杂的相互作用关系建模,重新分配各部分的权重,加大对动作细小差异的关注,提高对近似行为的分辨效果,最后将两路输出做分数融合,得到最终的判别结果。

本文剩余部分内容做如下安排:第 2 节介绍高阶注意力机制卷积神经网络;第 3 节介绍实验的相关设置、训练测试方法和结果分析;第 4 节为结论。

2　高阶注意力机制卷积神经网络

本文设计的人体行为识别方法整体流程如图 1 所示。

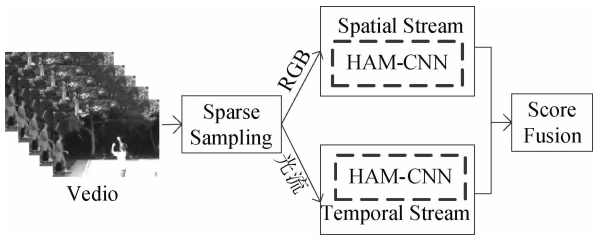


图 1　整体流程示意图

Fig. 1　Overview method flow

2.1　网络结构

本文在 TSN 的基础上进行了改进,采用的基础网络结构为 BN-Inception(Inception with Batch Nor-

malization),因为该网络可以兼顾准确率和效率之间的平衡。网络为 Inception 模块的堆叠,并引入批归一化层(Batch Normalization,BN)。Inception 模块将输入分为 4 个并行支路,分别采用 1×1、3×3、5×5 大小的滤波器进行卷积和池化,不同大小的卷积核获得不同大小的感受野,最后做不同尺度特征的融合,在增加网络宽度的同时,增加了网络对不同尺度的适应能力。模块中使用了多个 1×1 卷积,即在相同尺寸的感受野中叠加更多的卷积,能提取到更丰富的特征,同时,对输入特征图做降维,减少了计算的复杂度。BN-Inception 网络将 Inception 模块中 5×5 卷积用两个 3×3 卷积来替换,进一步降低计算量,如图 2 所示。批归一化层的使用,使每一层都规范化到一个 N(0,1)的高斯,避免网络层输出分布变化较大,梯度更新需要不断调整以适应输出分布变化带来的影响,使得收敛速度减慢。引入批归一化层,可以加速网络训练,较好的抑制过拟合。

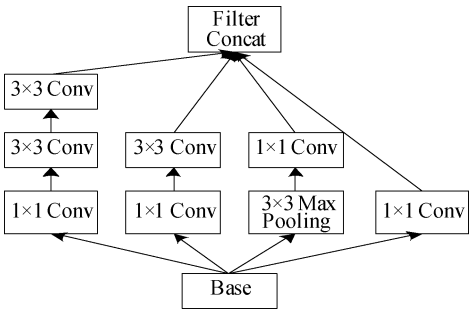


图 2 Inception 模块示意图

Fig. 2 Illustration of the inception module

在上述网络中引入高阶注意力模块(High-Order Attention Module, HAM),该模块放置于 inception(3c)层和 inception(4a)层之间,如图 3 所示。BN-Inception 网络共 33 层,该位置处于第 12 层之后。第一部分对输入图片从原始像素空间到中间层特征空间进行编码;第二部分编码注意力信息到能够分类的高层特征空间。高阶注意力模块的位置不宜太靠前,采集到的为低层次结构信息,包含有许多噪声;位置也不宜太靠后,一些具有辨别力的信息在前向传播过程中已经丢失。在全局平均池化层(Global Average Pooling,GAP)之后,加入全连接层(Fully Connected Layer,FC)和分类层,全连接层对网络中提取的有用信息进行整合,分类层的输出为数据集对应的动作类别数。

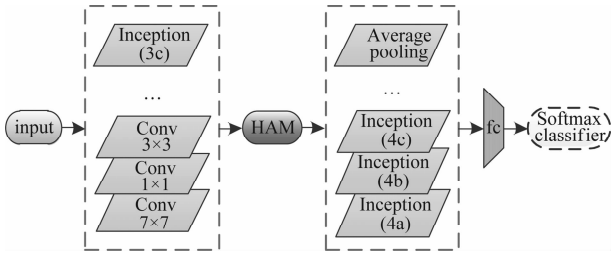


图 3 网络结构示意图

Fig. 3 Network architecture overview

2.2 高阶注意力模块

为了对各部分之间的复杂高阶相互作用进行建模,本文提出了高阶注意力模块,具体实现过程如下所述:首先定义一个线性多项式预测变量,$\boldsymbol{x}\in\mathbb{R}^C$ 为 $\boldsymbol{X}$ 的某个空间位置的局部描述符。

$$a(\boldsymbol{x})=\sum_{r=1}^{R}\langle \boldsymbol{w}^r,\otimes_r\boldsymbol{x}\rangle \tag{1}$$

其中,R 表示阶数,$\langle\cdot,\cdot\rangle$ 为两个相同尺寸张量的内积,$\otimes_r\boldsymbol{x}$ 为 $\boldsymbol{x}$ 的第 r 阶外积,其包括 $\boldsymbol{x}$ 中阶数的值都为 r 的单项式,$\boldsymbol{w}^r$ 为学习到的第 r 阶张量,它包括 $\boldsymbol{x}$ 中 r 阶变量组合的权重。

由于 $\boldsymbol{w}^r$ 较大的阶数 r 会引入大量的参数并且发生过拟合的问题,设当阶数 r 大于 1 时,$\boldsymbol{w}^r$ 通过张量的分解能够近似为 D^r 个一级张量,即,当 r 大于 1 时,$\boldsymbol{w}^r=\sum_{d=1}^{D^r}\alpha^{r,d}\boldsymbol{u}_1^{r,d}\otimes\cdots\otimes\boldsymbol{u}_r^{r,d}$,其中,$\boldsymbol{u}_1^{r,d}\in\mathbb{R}^C,\cdots,\boldsymbol{u}_r^{r,d}\in\mathbb{R}^C$ 是矢量,$\otimes$为外积,$\alpha^{r,d}$表示第 d 个一级张量的权重。根据张量的运算,公式(1)可以重新表达为:

$$\begin{aligned}a(\boldsymbol{x})&=\langle \boldsymbol{w}^1,\boldsymbol{x}\rangle+\sum_{r=2}^{R}\left\langle\sum_{d=1}^{D^r}\alpha^{r,d}\boldsymbol{u}_1^{r,d}\otimes\cdots\otimes\boldsymbol{u}_r^{r,d},\otimes_r\boldsymbol{x}\right\rangle=\\&\langle \boldsymbol{w}^1,\boldsymbol{x}\rangle+\sum_{r=2}^{R}\sum_{d=1}^{D^r}\alpha^{r,d}\prod_{s=1}^{r}\langle \boldsymbol{u}_s^{r,d},\boldsymbol{x}\rangle=\\&\langle \boldsymbol{w}^1,\boldsymbol{x}\rangle+\sum_{r=2}^{R}\langle \boldsymbol{\alpha}^r,\boldsymbol{z}^r\rangle\end{aligned} \tag{2}$$

上式中,$\boldsymbol{\alpha}^r=[\alpha^{r,1},\ldots,\alpha^{r,D^r}]^{\mathrm{T}}$ 表示权重矢量,$\boldsymbol{z}^r=[z^{r,1},\ldots,z^{r,D^r}]^{\mathrm{T}}$,其中,$z^{r,d}=\prod_{s=1}^{r}[\boldsymbol{u}_s^{r,d},\boldsymbol{x}]$。为了后面表达式的化简,式(2)可写为:

$$a(\boldsymbol{x})=1^{\mathrm{T}}(\boldsymbol{w}^1\odot\boldsymbol{x})+\sum_{r=2}^{R}(\boldsymbol{\alpha}^r\odot\boldsymbol{z}^r) \tag{3}$$

式中,$\odot$为哈达玛积,1^{T} 为一个 1 的行向量。为得到预测输出 $\boldsymbol{a}(\boldsymbol{x})\in\mathbb{R}^C$,通过引入辅助矩阵 $\boldsymbol{P}^r$ 来泛化表达式(3):

$$\boldsymbol{a}(\boldsymbol{x})=\boldsymbol{P}^{1\mathrm{T}}(\boldsymbol{w}^1\odot\boldsymbol{x})+\sum_{r=2}^{R}\boldsymbol{P}^{r\mathrm{T}}(\boldsymbol{\alpha}^r\odot\boldsymbol{z}^r) \tag{4}$$

式中，$\boldsymbol{P}^1\in\mathbb{R}^{C\times C}$，$\boldsymbol{P}^r\in\mathbb{R}^{D^r\times C}$，$r$ 为大于 1 的值。由于 $\boldsymbol{P}^r$，$\boldsymbol{w}^1$，$\boldsymbol{\alpha}^r$ 为要学习的参数，为了方便表达，由矩阵代数学，把 $\{\boldsymbol{P}^1,\boldsymbol{w}^1\}$ 转化为一个新的单矩阵 $\hat{\boldsymbol{w}}^1\in\mathbb{R}^{C\times C}$，同时，$\{\boldsymbol{P}^r,\boldsymbol{\alpha}^r\}$ 转化为 $\hat{\boldsymbol{\alpha}}^r\in\mathbb{R}^{D^r\times C}$。式(4)可以表达为：

$$\boldsymbol{a}(\boldsymbol{x})=\hat{\boldsymbol{w}}^{1\mathrm{T}}\boldsymbol{x}+\sum_{r=2}^{R}\hat{\boldsymbol{\alpha}}^{r\mathrm{T}}\boldsymbol{z}^r \tag{5}$$

上述等式包括两部分，将等式表达为更一般的情况，假设 $\hat{\boldsymbol{w}}^1$ 可以通过两个矩阵 $\hat{\boldsymbol{v}}\in\mathbb{R}^{C\times D^1}$ 和 $\hat{\boldsymbol{\alpha}}^1\in\mathbb{R}^{D^1\times C}$ 相乘来近似，即，$\hat{\boldsymbol{w}}^1=\hat{\boldsymbol{v}}\hat{\boldsymbol{\alpha}}^1$。等式(5)可以重新表达如下：

$$\boldsymbol{a}(\boldsymbol{x})=\hat{\boldsymbol{\alpha}}^{1\mathrm{T}}(\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{x})+\sum_{r=2}^{R}\hat{\boldsymbol{\alpha}}^{r\mathrm{T}}\boldsymbol{z}^r=\sum_{r=1}^{R}\hat{\boldsymbol{\alpha}}^{r\mathrm{T}}\boldsymbol{z}^r \tag{6}$$

上式中，$\boldsymbol{z}^1=\hat{\boldsymbol{v}}^{\mathrm{T}}\boldsymbol{x}$，当阶数 r 的值大于 1 时，$\boldsymbol{z}^r$ 与等式(2)中的值相同。$\hat{\boldsymbol{\alpha}}^r\in\mathbb{R}^{D^r\times C}$ 是能通过训练得到的参数。

在等式(6)中，$\boldsymbol{a}(\boldsymbol{x})$ 能够建模和使用局部描述符 $\boldsymbol{x}$ 的高阶统计信息，因此，我们可以在等式(6)上执行 sigmoid 函数来得到矢量形式的高阶注意力图：

$$\boldsymbol{A}(\boldsymbol{x})=\mathrm{sigmoid}(\boldsymbol{a}(\boldsymbol{x}))=\mathrm{sigmoid}\left(\sum_{r=1}^{R}\hat{\boldsymbol{\alpha}}^{r\mathrm{T}}\boldsymbol{z}^r\right) \tag{7}$$

式中，$\boldsymbol{A}(\boldsymbol{x})\in\mathbb{R}^C$，$\boldsymbol{A}(\boldsymbol{x})$ 中每个元素的值在[0,1]之间。

综上，$\boldsymbol{A}(\boldsymbol{x})$ 是在局部描述符 $\boldsymbol{x}$ 上定义的，为得到在 3D 张量 $\boldsymbol{X}$ 上定义的 $\boldsymbol{A}(\boldsymbol{X})$，推广等式(7)，在 $\boldsymbol{X}$ 的不同空间位置共享 $\boldsymbol{A}(\boldsymbol{x})$ 上学习到的权重，令 $\boldsymbol{A}(\boldsymbol{X})=\{\boldsymbol{A}(\boldsymbol{x}_{(1,1)}),\cdots,\boldsymbol{A}(\boldsymbol{x}_{(H,W)})\}$，其中，$\boldsymbol{x}_{(H,W)}$ 表示 $\boldsymbol{X}$ 上 (H,W) 点处的局部描述符。在卷积神经网络中使用注意力图有两个优点，一是不同位置权重共享，不会引入大量的参数；二是 $\boldsymbol{A}(\boldsymbol{X})$ 可以很容易通过 1×1 卷积实现。

$\boldsymbol{A}(\boldsymbol{X})$ 通过卷积操作实现，当 R 等于 1 时，矩阵 $\{\hat{\boldsymbol{v}},\hat{\boldsymbol{\alpha}}^1\}$ 通过输出分别为 D^1 和 C 通道数的 1×1 卷积层实现。当 R 大于 1 且 r 大于 1 时，首先将 $\{\boldsymbol{u}_s^{r,d}\}_{d=1,\cdots,D^r}$ 作为 D^r 个 1×1 卷积滤波器作用于 $\boldsymbol{X}$，来得到通道数为 D^r 的特征图 $\boldsymbol{Z}_s^r$，然后将特征图 $\{\boldsymbol{Z}_s^r\}_{s=1,\cdots,r}$ 做哈达玛乘积，得到 $\boldsymbol{Z}^r=\boldsymbol{Z}_1^r\odot\cdots\odot\boldsymbol{Z}_r^r$，其中，$\boldsymbol{Z}^r=\{\boldsymbol{z}^r\}$，$\hat{\boldsymbol{\alpha}}^r$ 也通过 1×1 卷积层实现。三阶注意力模块如图 4 所示。

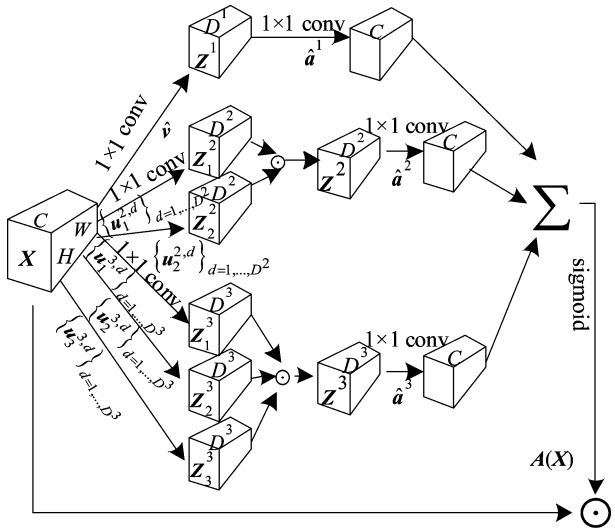


图 4　三阶注意力模块示意图

Fig. 4　Illustration of the third-order attention module

2.3　稀疏时间采样

传统双流卷积网络在空间网络输入单帧，时间网络输入短片段上一系列帧的堆叠，无法对长范围的视频内容进行学习。提出稀疏时间采样的方式，可以减少视频连续帧之间的冗余信息，同时可以大幅降低计算开销，即空间流卷积网络和时间流卷积网络以稀疏采样后的一系列短片段作为输入，以学习到整个视频的信息。具体实现为：将一段完整视频，用 V 表示，平均分为 N 段，用 $(V_1,V_2,\cdots,V_N)$ 表示，从上述分好的每个段中随机采样一个小片段 S_N，即 S_N 为 $(V_1,V_2,\cdots,V_N)$ 中对应的 V_N 段随机取出的一个小片段，每个小片段包含一帧 RGB 图片和两帧光流图片（x 方向和 y 方向），采样过程如图 5 所示。高阶注意力机制卷积神经网络可以用公式(8)来表示：

$$\begin{aligned}&\mathrm{HAM\text{-}CNN}(S_1,S_2,\cdots,S_N)=\\&M(g(F(S_1;P),F(S_2;P),\cdots,F(S_N;P)))\end{aligned} \tag{8}$$

式中，P 为网络的参数，$F(S_N;P)$ 为网络输出，即每个短片段对应的类别分数。g 为聚合函数，采用均值方式，即对不同片段在同一类别所得的分数值取平均，得到单一支路的判别结果（RGB 或光流）。将两支路的输出结果做分数融合，最后用 M 函数（softmax 函数）计算概率值，概率值最大的类别即视频所对应的动作分类。

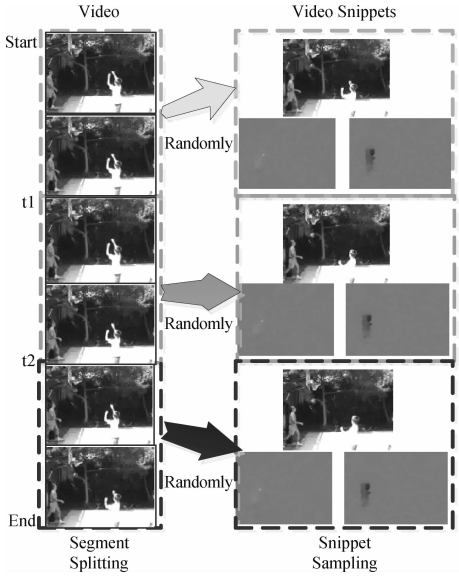


图 5 稀疏采样过程示意图

Fig. 5 Illustration of the progress of sparse sampling

3 分析与讨论

3.1 实验数据集与网络初始化

实验所用的硬件资源包括 Dell precision 7920 型服务器、两块显存为 11G 的 NVIDIA GTX 1080 TI 显卡,操作系统是 Ubuntu 16.04,使用的深度学习开发框架为 PyTorch 开源框架。

实验在两个主流视频数据集 UCF101 和 HMDB51 上进行,并在自制数据集 UCF24 上进行了对比实验。UCF101 数据集包含 101 类动作,每类动作由 25 个人完成,每人做 4 ~ 7 组,总计 13320 个视频,动作可以分为 5 大类,即人与物体交互、人体动作、人与人交互、乐器演奏、体育运动。HMDB51 数据集包含 51 类动作,每类动作至少包含 101 个视频,共 6849 个视频,多来源于大量现实视频的集合,如电影和网络视频等的剪辑片段。UCF24 为从 UCF101 中提取部分视频自制而成,该数据集包含 24 类动作,分为 6 大类近似行为,共 3188 个视频。如吊环、双杠、鞍马、高低杠为一大类近似动作。部分动作类别如图 6 所示。本文按照文献[30,31]中对数据集的划分方法,将 UCF101 和 HMDB51 这两个数据集分别划分为三个子集(split),最终的识别结果为三个分割子集所求得的结果取线性平均。

相比于所设计神经网络含有的参数量,本文的数据集数据量很小,若仅采用人体行为识别数据集对网络做训练,网络会出现过拟合的现象。故采用迁移学习的方法,将 ImageNet 数据集在 BN-Inception 网络上预训练得到的参数权重,作为本文设计网络的初始权重,然后用人体行为数据集对设计网络微调。具体初始化方式为:BN-Inception 网络对应层参数采用 ImageNet 数据集训练得到的权重初始化;全连接层采用 Kaiming 初始化;分类层采用均值为 0,标准差为 0.001 的高斯分布随机初始化。

图 6 UCF24 部分动作类别示意图

Fig. 6 Illustration of the part of the action category of UCF24

3.2 网络训练与测试

训练时,训练视频占视频总量的 7/10,将视频等分为 3 段,采用随机梯度下降法(stochastic gradient descent,SGD)来学习模型参数,batch_size 设置为 16,动量和权重衰减因子依照经验设置,分别为 0.9 和 0.0005。初始学习率设置为 0.001,对空间网络,迭代 80 回合,每 30 回合学习率下降为原来的 1/10;对时间网络,迭代 340 回合,分别在 190 回合和 300 回合下降为原来的 1/10。

训练过程中采用如下三个训练策略来避免过拟合现象的发生:

(1)交叉输入模式预训练,由于 RGB 与光流分布不同,所以时空网络采用不同的初始化方式。ImageNet 为图片数据集,因此,空间网络采用在 ImageNet 上训练的模型做初始化,对于时间网络,将光流场的范围线性离散到与 RGB 相同的 0 ~ 255 区间,对第一个卷积层的权重在 RGB 通道数上取平均,将均值乘以时间网络输入的通道数,作为时间网络第一个卷积层的权重,来达到降低时间网络过拟合的目的。

(2)正则化技术,即部分的批归一化。由于批归一化通过有限的训练样本来调节输出的分布容易造成过拟合,所以,初始化后,除了第一个卷积层,对其他所有的 BN 层参数冻结。第一个卷积层的激活值对 RGB 和光流重新做估计。

(3)数据扩充,稀疏采样得到的图片尺寸为 320 ×240,对图片从边角、中心提取区域,得到 224×224

大小的图片；对图片的宽高从{256,224,192,168}中随机选择两个数裁剪，裁剪后的区域尺寸重新设置为 224×224 大小。通过上述两种方式，达到增加训练样本数量和多样性的目的。

测试时，测试视频占视频总量的 3/10，从动作视频中采样 25 个 RGB 帧或光流堆，将采样图片裁剪4 个边角和 1 个中心，并做水平翻转，将所得图片送入网络测试。空间流和时间流网络采用加权平均的方式进行融合，空间流与时间流的权重设置为 1∶1.5。在 softmax 归一化之前融合 25 帧和不同流的预测分数。

3.3　损失函数

网络采用交叉熵损失函数，该函数常用于解决多分类问题，由前向传播过程得到，用来判定实际输出与预测输出的接近程度，差值越小表明预测效果越好。将所得误差值做反向传播来进行参数更新。具体表示如下：

$$L(y,\boldsymbol{G})=-\sum_{i=1}^{C} y_i\left(G_i-\log\sum_{j=1}^{C}\exp G_j\right) \tag{9}$$

C 为类别数，y_i 为类别 i 对应的真实值。$G_i=g(F_i(S_1;P),F_i(S_2;P),\cdots,F_i(S_N;P))$，指所有片段判断为类别 i 的得分采用 g 函数来得到对应视频类别 i 的分数。$\boldsymbol{G}$ 是一个长度为 C 的向量，每个元素代表视频属于该类别的分数。

网络是可微的，在反向传播过程中，参数 P 通过多个分割片段的联合来优化，损失值 L 对模型参数 P 的梯度为：

$$\frac{\partial L(y,\boldsymbol{G})}{\partial P}=\frac{\partial L}{\partial G}\sum_{n=1}^{N}\frac{\partial G}{\partial F(S_n)}\frac{\partial F(S_n)}{\partial P} \tag{10}$$

损失值通过迭代更新模型参数来优化，模型中参数的学习不靠单一的短片段，而是从整个视频范围来学习，可以更好建模长时间结构的视频信息。

3.4　识别结果与分析

实验探究了不同阶数注意力模块对卷积网络识别性能的影响，在 TSN 网络中引入高阶注意力模块，在 UCF101 数据集上以 RGB 形式作为输入，实验结果如图 7 所示，由图可知，随着注意力模块阶数的增加，与 TSN[16] 的识别结果 85.1% 相比，识别准确率均取得了一定的提升，其中四阶注意力机制卷积网络提高了 1.53%，五阶和六阶网络分别提高了 0.19% 和 0.20%，提升幅度较小，分析为随着阶数的提高，建模到特征图中一些次要的边缘信息，影响了识别的准确率。

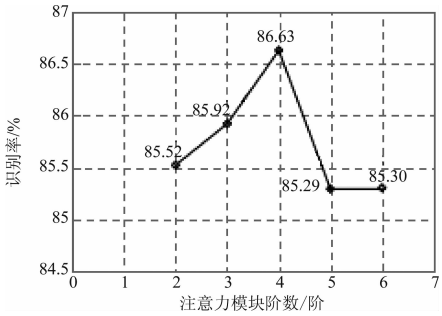


图 7　不同阶数注意力模块对识别性能的影响

Fig. 7　Effects of different order attention modules on recognition performance

由上述实验可得，四阶注意力模块可以较好建模动作部分之间的相互作用关系，提高行为识别的准确率，所以，后续实验均采用四阶注意力机制卷积神经网络。

笔者探究了高阶注意力模块在网络中所处位置对识别性能的影响，如下表 1 所示，由表中实验结果可知，高阶注意力模块放置于 inception(3c)层之后，识别结果取得了较明显的提升，而位置靠前或靠后，识别效果均不理想，分析为靠前结构信息较粗糙，靠后则部分信息在前向传播过程中已经丢失，因此，后续实验均采用模块置于 inception(3c)层之后进行。

表 1　注意力模块处于不同位置对识别性能的影响

Tab. 1　The effect of different position of attention module on recognition performance　%

UCF101(RGB)				
方法	Split1	Split2	Split3	准确率
TSN[16]	85.5	84.9	84.5	85.1
HAM-inception(3a)	84.83	84.15	84.58	84.52
HAM-inception(3c)	87.75	85.79	86.35	**86.63**
HAM-inception(4e)	84.56	85.50	85.39	85.15

* HAM-inception(3a)即高阶注意力模块放置于 inception(3a)层之后，其他依此类推。

为验证高阶注意力机制对近似行为的识别效果，在自制数据集 UCF24 上进行了对比实验。实验结果如下表 2 所示，由表中数据可得，网络中引入高阶注意力机制后，识别率提高了 1.7%，较原始网络取得了较明显的提升，验证了该方法可以关注到动作之间的微小差别，对近似行为具有较好的识别效果。

表 2　不同方法在 UCF24 数据集上的识别率对比

Tab. 2　Compare the recognition rates of different methods on UCF24 dataset　%

UCF24			
方法	RGB	光流	准确率
TSN*	84.7	90.5	93.6
HAM-CNN(ours)	88.3	92.1	**95.3**

在UCF101数据集上进行了相应的实验,如表3所示,实验结果表明,与现有方法相比,采用本文提出的方法,识别准确率达到了94.7%,比TSN提高了0.7%,优于现有方法。

表4为各种方法在HMDB51数据集上的识别率对比,由实验结果可知,采用引入高阶注意力机制的人体行为识别方法,识别准确率较TSN提高了1.1%,达到了69.6%。较目前最优方法还存在一定差距,分析为该数据集取材于现实场景,行为之间背景差异较大,近似行为数量较少,不符合高阶注意力机制关注动作细粒度信息的特性,限制了该方法识别效果的进一步提升。

表3 不同方法在UCF101数据集上的识别率对比

Tab.3 Compare the recognition rates of different methods on UCF101 dataset %

UCF101	
方法	准确率
Two Stream[11]	88.0
C3D(3 nets)[17]	85.2
DMC-Net[32]	90.9
TSN[16]	94.0
Spatio-temporal Network[33]	94.6
ARTNet[34]	94.3
HAM-CNN(ours)	**94.7**

表4 不同方法在HMDB51数据集上的识别率对比

Tab.4 Compare the recognition rates of different methods on HMDB51 dataset %

HMDB51	
方法	准确率
Two Stream[11]	59.4
C3D(3 nets)[17]	51.6
DMC-Net[32]	62.8
TSN[16]	68.5
Spatio-temporal Network[33]	68.9
ARTNet[34]	**70.9**
HAM-CNN(ours)	69.6

在上述两个数据集上的实验结果表明,与现有方法相比,本文提出的方法表现出较好的性能,一定程度上提高了视频行为识别的准确率,原因在于可以较好建模动作部分之间的相互作用关系对最终判定结果的影响,提高对类间差异较小的行为间的分辨能力。

4 结论

本文提出一种引入高阶注意力机制的人体行为识别方法,在TSN中引入高阶注意力模块,建模特征图各个部分之间的相互作用关系,产生区别性的关注建议,捕获行为之间的细微差异。在UCF101和HMDB51这两个数据集上的实验结果表明,与TSN相比,识别的准确率分别取得了0.7%和1.1%的提升,验证了所提出方法的有效性。不足之处在于,行为类别的判定为采用各个段分数取平均的方式,视频的边缘区域中存在较大的噪声,影响了识别的精度。今后工作中,将尝试在高阶注意力机制卷积神经网络中引入一种新的段分数融合方式,进一步提高动作识别的准确率和鲁棒性。

参考文献

[1] 罗会兰, 童康, 孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报, 2019, 47(5): 1162-1173.
Luo Huilan, Tong Kang, Kong Fansheng. The Progress of Action Recognition in Videos Based on Deep Learning: A Review[J]. Acta Electronica Sinica, 2019, 47(5): 1162-1173. (in Chinese)

[2] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001(3): 257-267.

[3] Bobick A, Davis J. An appearance-based representation of action[C]//Proceedings of 13th International Conference on Pattern Recognition. IEEE, 1996, 1: 307-312.

[4] Weinland D, Ronfard R, Boyer E. Free viewpoint action recognition using motion history volumes[J]. Computer Vision and Image Understanding, 2006, 104(2-3): 249-257.

[5] Yilmaz A, Shah M. Actions sketch: a novel action representation[A]. Computer Vision and Pattern Recognition[C]//USA: IEEE, 2005: 984-989.

[6] Yang X, Tian Y L. Effective 3d action recognition using eigenjoints[J]. Journal of Visual Communication and Image Representation, 2014, 25(1): 2-11.

[7] Wang H, Klaser A, Schmid C, et al. Action Recognition by Dense Trajectories[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011.

[8] 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). IEEE, 2005, 1: 886-893.

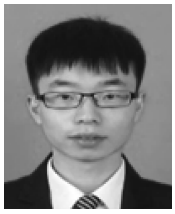
[9] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.

[10] Wang H, Schmid C. Action recognition with improved trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision, 2013: 3551-3558.

[11] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems, 2014: 568-576.

[12] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1933-1941.
[13] Zhang B, Wang L, Wang Z, et al. Real-time action recognition with enhanced motion vector CNNs[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2718-2726.
[14] Wang L, Xiong Y, Wang Z, et al. Towards good practices for very deep two-stream convnets[J]. ArXiv preprint arXiv: 1507.02159, 2015.
[15] Feichtenhofer C, Pinz A, Wildes R P. Spatiotemporal residual networks for video action recognition. CoRR abs/1611.02155 (2016)[J]. ArXiv preprint arXiv: 1611.02155, 2016.
[16] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016: 20-36.
[17] 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.
[18] Diba A, Fayyaz M, Sharma V, et al. Temporal 3d convnets: New architecture and transfer learning for video classification[J]. ArXiv preprint arXiv: 1711.08200, 2017.
[19] Tran D, Ray J, Shou Z, et al. Convnet architecture search for spatiotemporal feature learning[J]. ArXiv preprint arXiv: 1708.05038, 2017.
[20] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 5533-5541.
[21] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.
[22] Mnih V, Heess N, Graves A. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems, 2014: 2204-2212.
[23] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[C]// Neural Information Processing Systems, 2015: 2017-2025.
[24] Hu J, Shen L, Sun G, et al. Squeeze-and-Excitation Networks[C]// Computer Vision and Pattern Recognition, 2018: 7132-7141.
[25] Yang Y, Deng C, Gao S, et al. Discriminative Multi-instance Multitask Learning for 3D Action Recognition[J]. IEEE Transactions on Multimedia, 2017, 19(3):519-529.
[26] Yang Y, Deng C, Tao D, et al. Latent max-margin multitask learning with skelets for 3-D action recognition[J]. IEEE Transactions on Cybernetics, 2016, 47: 439-448.
[27] Xie D, Deng C, Wang H, et al. Semantic Adversarial Network with Multi-scale Pyramid Attention for Video Classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 9030-9037.
[28] Yang Y, Liu R, Deng C, et al. Multi-task human action recognition via exploring super-category[J]. Signal Processing, 2016, 124:36-44.
[29] Chen B, Deng W, Hu J. Mixed high-order attention network for person re-identification[C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 371-381.
[30] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. ArXiv preprint arXiv: 1212.0402, 2012.
[31] Kuehne H, Jhuang H, Garrote E, et al. HMDB: a large video database for human motion recognition[C]//2011 International Conference on Computer Vision. IEEE, 2011: 2556-2563.
[32] Shou Z, Lin X, Kalantidis Y, et al. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 1268-1277.
[33] Wang Y, Long M, Wang J, et al. Spatiotemporal pyramid network for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1529-1538.
[34] Wang L, Li W, Li W, et al. Appearance-and-relation networks for video classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1430-1439.

作者简介



王增强　男，1994 年生，山西吕梁人。中国民航大学电子信息与自动化学院硕士研究生，主要研究方向为计算机视觉与视频行为识别。
E-mail: 13622101396@163.com

张文强　男，1997 年生，山西朔州人。中国民航大学电子信息与自动化学院硕士研究生，主要研究方向为深度学习与人体动作识别。
E-mail: 745150158@qq.com

张　良(通信作者)　男，1970 年生，山东淄博人。中国民航大学电子信息与自动化学院教授，博士，主要研究方向为图像处理、模式识别、计算机视觉。
E-mail: l-zhang@cauc.edu.cn