DOI: 10. 13873 / J. 1000-9787(2020) 10-0035-04

基于深度学习的时空特征融合人体动作识别*

王　倩, 孙宪坤, 范冬艳
(上海工程技术大学 电子电气工程学院, 上海 201620)

摘　要: 深度学习需要充分利用视频中动作的时空信息来进行动作识别。为了充分利用视频中的时空特征来提高动作识别的准确率,并以较低的成本保存相关信息,提出一种采用稀疏采样方案的时空特征融合动作识别框架。采用稀疏采样获得视频的 RGB 图和光流图,分别送入 VGG-16 网络提取视频的时空特征;融合时空卷积神经网络(CNN) 提取中层时空融合特征;将中层时空融合特征送入 C3D CNN 识别出动作的类别。在 HMDB51 和 UCF101 两个数据集的实验结果表明: 该框架能够充分利用视频的时间信息和空间信息,达到了较高的动作识别准确率。

关键词: 深度学习; 动作识别; 稀疏采样; 时空特征融合; C3D 卷积神经网络(CNN)
中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 1000-9787(2020) 10-0035-04

Fusion of spatio-temporal features based on deep learning for human action recognition*

WANG Qian, SUN Xiankun, FAN Dongyan
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Deep learning needs to make full use of the spatio-temporal information of the actions in the video to perform action recognition. In order to make full use of the spatio-temporal features in video to improve the accuracy of action recognition and save relevant information at a lower cost, a spatio-temporal feature fusion action recognition framework using sparse sampling scheme is proposed. The framework uses the sparse sampling to obtain the RGB images and optical flow images of videos, and respectively sends the spatio-temporal features to the VGG-16 network to extract the spatio-temporal features, then the spatial CNN and the temporal CNN are merged to extract the fused spatio-temporal features middle level, finally the fused spatio-temporal features are sent to the C3D CNN to performs action recognition. The experimental results of two datasets in HMDB51 and UCF101 show that the framework can make full use of the temporal information and spatial information of the video to achieve higher action recognition accuracy.

Keywords: deep learning; action recognition; sparse sampling; spatio-temporal feature fusion; C3D convolutional neural network(CNN)

0　引　言

深度学习被广泛应用于计算机视觉领域,在物体、场景和复杂事件的图像分类方面取得了巨大成功[1,2]。对于视频的动作识别[3~7],在安全和行为分析等领域有重要的应用。然而不同于图像的是视频有两个关键和互补的方面: 外观和动态,因此端到端的深层网络仍然无法比基于传统手工特征的视频动作识别获得显著的优势。为了提取利用视频的时间特征和空间特征来进行动作识别,过去主要采用以下两种方法:

1) 采用并行的双流卷积神经网络(convolutional neural network, CNN) 分别提取视频的时间特征和空间特征。文献[3]首先提出了时空双流深度学习策略,使用两个独立的 CNN 中学习视频的空间特征和时间特征,分别进行视频行为动作的识别,再将 SoftMax 层的分数通过晚期融合(late fusion) [4] 进行合并。然而连续的帧是高度冗余的,对于高度相似帧的密集时间采样是不必要的,因此文献[7]提出了 TSN(temporal segment networks) 视频级框架。该框架利用稀疏采样方案在长视频序列上提取短片段,采用分段结构来聚合来自采样片段的信息,从而在合理的时间和计算资源预算下实现长时间视频序列的端到端学习。双流

收稿日期: 2019-06-26
* **基金项目**: 国家自然科学基金青年科学基金资助项目(61802251, 61801286) ; 上海市科学技术委员会科研计划项目(16DZ1206000) ; 上海工程技术大学科研项目(E3-0903-19-01053)

CNN 单独提取时间信息和空间信息,仅在最后进行时空特征的结合,没有考虑到时间信息和空间信息的关联性来充分融合时空信息,并且时空神经网络的全连接层在某些程度上已经破坏了视频的时间和空间特征,识别准确率仍然有待提高。

2) 采用三维(3D) CNN 来同时提取视频的时间信息和空间信息。文献[8]首次提出 3D CNN 结构,该结构是原先 2D 神经网络在时间维度上的一种扩展,使得网络可以学习视频片段时间上的特征。文献[5]提出的 C3D 神经网络结构使用若干个连续的视频帧作为输入,使用大小为 3 ×3 ×3 卷积核学习视频的时空特征,在卷积层和池化层均进行 3D 操作,提高了动作识别的准确率。但是直接使用 3D CNN 进行动作识别时,现有的网络还不能做到充分的利用视频的时空特征,因此识别准确率较差。

为了充分融合视频中的时空信息来进行动作识别,并以相当低的成本保存相关信息,本文在 TSN 网络的基础上,结合 C3D CNN,提出了时空特征融合动作识别模型(spatio-temporal feature fusion action recognition model,STFF-C3D CNN) 。

1 稀疏采样时空双流视频动作识别模型设计

1.1 整体架构设计

本文提出的 STFF-C3D CNN 框架主要分为四部分: 稀疏采样生成 RGB 图和光流图、时空特征的提取、时空混合特征图的生成、C3D CNN 进行动作识别,如图 1 所示。

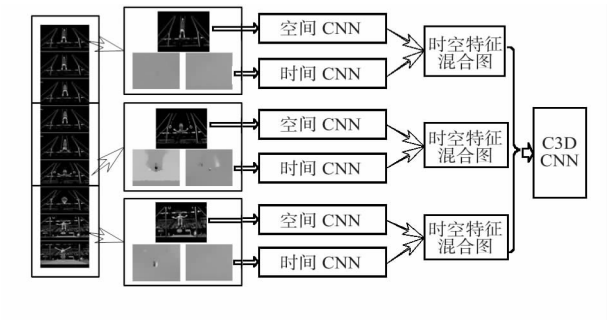

图 1 STFF-C3D CNN 框架

首先,采用稀疏采样方案对视频进行采样。稀疏采样不是处理单个帧或帧堆栈,而是对整个视频中稀疏采样的一小段短片段进行操作。一个输入视频被分为 K 段(segment) ,一个片段(snippet) 从它对应的段中随机采样得到。在形式上,给定一个视频 V,把它分成相等持续时间的 K 段 $\{S_1,S_2,\cdots,S_k\}$,$\{T_1,T_2,\cdots,T_k\}$ 是 snippets 的序列。每个 snippets 的 T_k 是从其相应的片段 S_k 随机采样得到的。对于每段 snippet,提取它包含空间信息的 RGB 图像和包含时间信息的 x 方向光流图和 y 方向光流图。

其次,训练空间 CNN 和时间 CNN 提取时间和空间特征。用 RGB 图像作为空间 CNN 的输入进行训练,提取空间中层特征,用光流图作为时间 CNN 的输入进行训练,提取时间中层特征。

然后,训练时空混合 CNN 提取时空融合特征。将时空中层特征图进行融合[9] ,训练混合 CNN,提取混合 CNN 的中层混合特征,生成时空混合特征图。可不破坏时空特征且能够将时空特征融合从而提取时空相关性信息。

最后,将时空混合特征图作为 C3D CNN 的输入,C3DCNN 在时间和空间维度分别进行卷积和池化,经过最后的 SoftMax 层进行动作识别,来同时学习视频的运动信息和静态的图片信息。

1.2 时空双流流 CNN

空间流 CNN 的输入是将视频稀疏采样后产生的一系列单帧 RGB 图像。RGB 图像使用三个通道存储像素信息,其中包含了视频的外形信息,外形信息是动作识别的重要信息。由于视频的动作识别和某些物体有密切的关联性,RGB 图像可以提取视频的空间特征,因此仅通过一系列的 RGB 图像作为 CNN 的输入也可以做视频的动作识别。

时间流 CNN 的输入是一系列的视频帧之间的光流图像,本文输入的光流图像是稀疏采样得到的 x 方向光流图和 y 方向光流图。光流场是指图像中所有像素点构成的一种二维瞬时速度场,其中的二维速度矢量是景物中可见点的三维速度矢量在成像表面的投影。所以光流不仅包含了被观察物体的运动信息,而且还包含有关景物三维结构的丰富信息[10] 。把光流图作为神经网络的输入可以有效提取视频动作的时间特征,提高视频人体动作识别的准确率。

本文采用的时间 CNN 和空间 CNN 是首次出现在 2014 年 ImageNet 图像分类竞赛中获得亚军的 VGG16 网络[1] 。其网络结构图如图 2 所示。

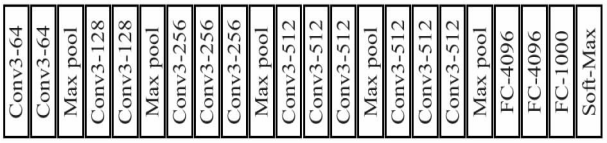

图 2 VGG16 网络结构

1.3 时空特征融合

时空网络的融合能合理利用视频的空间特征和时间特征,并结合二者的关联性来判别动作的类型。例如,在撑杆跳行为中,空间流网络识别出了静态的人体和杆子,时间流网络识别出了在一定的空间位置人体进行翻转运动。同样在化眼妆动作中,空间流卷积网络识别出了静态的人脸和手,时间流网络识别出了在一定的空间位置手部进行周期性的动作。因此结合时空 CNN 可以辨别出撑杆跳和化眼妆这两个动作类型。这里主要阐述 STFF-C3D CNN 框架的时空特征融合方法。

由于全连接层在某些程度上已经破坏了视频的时间和空间特征,为了利用空间与时间特征在像素级别上的相关

性,本文选择了在卷积层进行了融合,具体结构如图 3 所示。该结构在第九层(第七层卷积层) 将空间流 CNN 融合到时间流中,并且去除了空间流 CNN 在第九层之后的结构。之后在训练集上再次进行训练,通过前馈与方向传播调整融合后的结构参数。最后,如图 4 所示,运用训练好的基于特征融合的双流卷积神经网络,将需要分类的视频输入到网络当中,输出混合卷积层的特征图作为人物行为动作的中层特征。

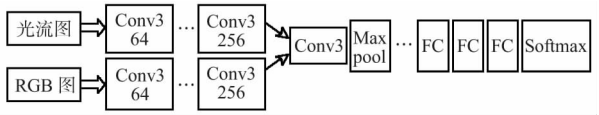

图 3　时空双流融合网络结构

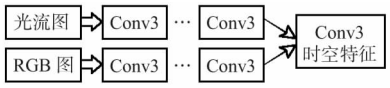

图 4　时空特征提取模型结构

时空 CNN 的融合位置可以在 VGG CNN 的任何位置,但是神经网络之间的融合跟特征图的大小和神经网络的通道数有关系,主要考虑以下两点约束: 1) 特征图的大小需要调整到一致大小,若大小不一致将较小的特征图做上采样操作; 2) 空间流 CNN 与时间流 CNN 通道数量应当一致。

本文所使用的时空特征融合策略可表述为

$$F^{\mathrm{sum}} = f^{\mathrm{sum}}(p^a, p^b) \tag{1}$$

$$F^{\mathrm{sum}}_{i,j,d} = p^a_{i,j,d} + p^b_{i,j,d} \tag{2}$$

式(1) 表示将两个网络的特征图 $p^a \in R^{H\times W\times D}$ 和 $p^b \in R^{H'\times W'\times D'}$ 通过求和的方式将两个网络的特征融合成一个新的特征图 $F^{sum} \in R^{H''\times W''\times D''}$,其中,$H$ 表示特征图的高度,W 表示特征图的宽度,D 表示特征图通道数,并且满足关系 $H = H' = H''$,$W = W' = W''$,$D = D' = D''$。该公式能被应用于卷积层、全连接层及池化层的融合。

式(2) 具体描述了如何使用求和的方法在第 d 通道特征图的像素点(i,j) 处进行融合,其中 $1 \leqslant i \leqslant H$,$1 \leqslant j \leqslant W$,$1 \leqslant d \leqslant D$,$p^a$,$p^b$,$F \in R^{H\times W\times D}$。

1.4　基于 C3D CNN 的动作识别

C3D CNN 与 2D CNN 不同,它有选择地兼顾运动和外观。在撑杆跳例子中,特征先是集中在整个人身上,然后跟踪其余帧上的撑杆跳表演的动作。同样在化眼妆例子中,它先聚焦在眼睛上,然后在化妆时跟踪眼睛周围发生的动作。C3D CNN 不仅对空间的水平和竖直维度进行卷积和池化,对时间维度也进行了卷积和池化,以更好地提取时间和空间特征,保持时空特征的相关性。

本文采用的 C3D CNN 结构如图 5 所示。共具有 8 个卷积层、5 个池化层、2 个全连接层,以及一个 SoftMax 输出层。所有 3D 卷积滤波器均为 3 × 3 × 3,步长为1 × 1 × 1。为了保留早期的时间信息,pool1 核大小为 1 × 2 × 2、步长 1 × 2 × 2,其余所有 3D 池化层均为 2 × 2 × 2,步长为2 × 2 × 2。每个全连接层有 4 096 个输出单元。C3D CNN 的第一层输入是由前面双流 CNN 经过特征融合之后的中层特征图 $M \in R^{H\times W\times D\times T}$,其中,$H$,$W$ 和 D 分别是时空特征融合图的高度、宽度以及通道数,T 是输入的特征融合图的数量。C3D 最终通过 SoftMax 层给出视频样本的分类。

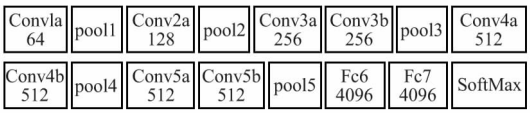

图 5　C3D 网络结构

2　实　验

2.1　实验设计

为了对本文所提出的框架进行有效的评估,本实验在两个著名的视频动作识别数据集 HMDB51 和 UCF101 上进行实验,视频均为真实场景下拍摄,充分考虑个场景类型、视角区域以及光照、背景变化与摄像头移动。UCF101 数据集共有 101 类人体行为,共 13 320 段不同视频,总时长 27 h,每类视频有 25 组,每组视频序列由 4 ~ 7 段视频段组成,每段视频段从 1.06 ~ 71.04 s 不等。HMDB51 数据集[11] 中的视频多数来源于电影,还有一部分来自公共数据库以及 YouTube 等网络视频库。由来自 51 个动作类别的 6 766 个视频剪辑组成,每类至少包含有 101 段样本。

本文的 STFF-C3D CNN 实验分为三个步骤: 1) 采用稀疏采样方法预训练时间 CNN 和空间 CNN; 2) 训练混合 CNN 提取时空融合特征; 3) C3D CNN 进行动作识别。

1) 预训练时空 CNN: 网络结构采用 VGG-16[1] 结构,用稀疏采样对视频进行采样。对于空间 CNN,使用 RGB 图像作为输入。对于时间 CNN,由于输入的是 x 方向光流图和 y 方向光流图,还要对网络结构做修改,主要差别在第一个卷积层,因该层的输入通道依据不同的输入类型而变化。除第一个卷积层外,其它层通道数和网络设置相同。时空 CNN 的输入均为动作类别。为了更充分地提取特征,将所有视频帧采用尺寸改为 224 × 224,并通过数据增强将一张图片产生 10 张图片。即将每张图片的 4 个点的坐标和中心坐标当做左上角坐标裁剪出指定尺寸的 5 张图片,再对裁剪得到的图像做左右翻转,最后把未翻转的和翻转后的列表合并,这样一张输入图像可得到 10 张输出。为了证明仅利用 RGB 图像或光流图也能实现动作识别,实验对训练好的时空 CNN 分别进行了测试。

2) 训练混合 CNN 提取时空融合特征: 全连接层已经一定程度上破坏了某些时间和空间特征,因此实验在第九层(第七层卷积层) 将空间流 CNN 融合到时间流中。没有在第九层之后进行卷积的原因是,特征图经过两层池化层之后大小缩小到了 56 × 56,若经过多池化层,会对时空特征造

成一定的破坏并且过度缩小特征图。训练时,通过不断反向传播对融合后的参数进行调整。训练结束后,混合 CNN 被用作提取混合时空特征。为了更好地理解时空特征,实验将提取出来的时空特征进行了可视化。图 6(d)~(f) 分别为 RGB 图、水平和竖直方向的光流图经过 VGG-16 网络后的中层特征可视化图,按顺序分别为第二层、第五层、第九层的中层特征。

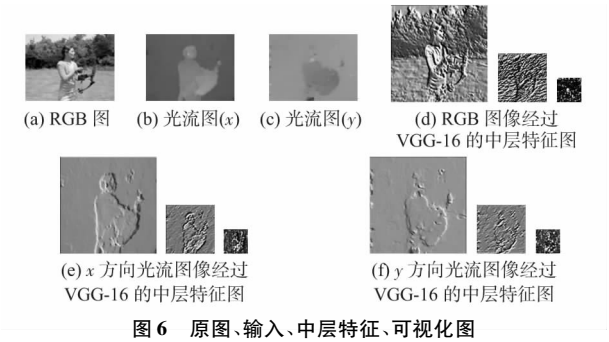

图 6 原图、输入、中层特征、可视化图

3) C3D CNN 进行动作识别: 将提取的时空融合特征图作为输入训练 C3D CNN 的。这一步中 C3D CNN 将利用三维卷积和三维池化进一步提取视频的时空特征,完成视频的动作识别任务。

2.2 结果对比分析

实验在 Ubuntu 16.04 系统上进行,服务器使用 2 块 GPU,显卡型号是 GTX1080,内存大小为 16G,在 Pytorch 深度学习框架下实现。实验将数据集平均分成 3 份进行训练和测试,将 3 次测试结果的平均值作为最后的准确率。

如表 1 所示,实验只将 RGB 图或者光流图输入 TSN[7] 网络进行训练和测试,通过数据可知仅通过 RGB 图像(RGB-TSN) 也可以识别出动作类型,仅通过光流图(flow-TSN) 可以比仅通过 RGB 图获得更高的准确率; 本文框架比文献[9]的 Spatiotemporal-3D CNN 框架的准确率高,算法复杂度更低,训练效率更高; 实验证明在都使用 VGG-16 作为时空 CNN 的情况下本文框架比 TSN 框架获得更高的准确率; 本文的框架比使用 BNInception 作为时空 CNN 的 TSN 框架的准确率也得到了提高。实验表明,本文提出方法能够有效的识别动作类别。

表 1 本文动作识别算法和其他算法准确率对比 %

动作识别算法	UCF101 准确率	HMDB51 准确率
Two-stream[3]	88.0	59.4
C3D[5]	85.2	–
RGB-TSN	86.4	56.4
Flow-TSN	87.7	64.2
Spatiotemporal-3D CNN[9]	91.7	65.6
TSN(BNInception) [7]	94.0	68.5
TSN(VGG-16)	92.5	66.8
STFF-C3D CNN(VGG-16)	94.8	69.5

3 结 论

为了充分利用视频中的时空特征来提高动作识别的准确率,并以较低的成本保存相关信息,本文提出一种采用稀疏采样方案的时空特征融合动作识别框架。在 UCF101 和 HMDB51 数据集上进行实验和对比,实验结果表明: 该模型能够有效识别视频中的动作类型,达到了较高的准确率。

参考文献:

[1] SIMONYAN K, ZISSERMAN A. Very deep convol-utional networks for large-scale image recognition [C]// Proceedings of International Conference on Learning Representations, 2015: 1 – 14.

[2] XIONG Y, ZHU K, LIN D, et al. Recognize complex events from static images by fusing deep channels [C]// Computer Vision & Pattern Recognition, IEEE, 2015.

[3] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1(4) : 568 – 576.

[4] KARPATHY A, TODERICI G, SHETTY S, et al. Large scale video classification with convolutional neural networks [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , IEEE Computer Society, 2014: 1725 – 1732.

[5] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]// Proc of IEEE International Conference on Computer Vision, Piscataway, NJ: IEEE Press, 2015: 4489 – 4497.

[6] 胡珂杰,蒋敏,孔军. 基于混合关节特征的人体行为识别[J]. 传感器与微系统,2018,37(3) : 138 – 144.

[7] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of European Conference on Computer Vision, Berlin: Springer, 2016: 20 – 36.

[8] JI S, XU W, YANG M, et al. 3D Convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1) : 221 – 231.

[9] 杨天明,陈志,岳文静. 基于视频深度学习的时空双流人物动作识别模型[J]. 计算机应用,2018,38(3) ,895 – 899.

[10] 李崇国. 光流法在靶区运动图像配准中的应用研究[D]. 泸州: 泸州医学院,2009.

[11] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A large video database for human motion recognition [C]// IEEE International Conference on Computer Vision, 2011: 2556 – 2563.

作者简介:

王 倩(1995 –) ,女,硕士研究生,研究方向为计算机视觉与图形处理,E-mail: wq_wzz_ok@163.com。

孙宪坤(1972 –) ,男,博士研究生,副教授,主要研究领域为计算机视觉、计算机控制及其应用。