文章编号:1671-4598(2018)08-0266-04 DOI:10.16526/j.cnki.11—4762/tp.2018.08.059 中图分类号:TP242 文献标识码:A

基于双流卷积神经网络的改进人体行为识别算法

张怡佳，茅耀斌
(南京理工大学 自动化学院，南京 210094)

摘要：近年来人体行为识别成为计算机视觉领域的一个研究热点，而卷积神经网络（convolutional neural network，CNN）在图像分类和识别领域取得了重要突破，但是人体行为识别是基于视频分析的，视频包含空间域和时间域两部分的信息；针对基于视频的人体行为识别问题，提出一种改进的双流卷积神经网络（Two—Stream CNN）模型，对于空间域，将视频的单帧 RGB 图像作为输入，送入 VGGNet _ 16 模型；对于时间域，将多帧叠加后的光流图像作为输入，送入 Flow _ Net 模型；最终将两个模型的 Softmax 输出加权融合作为输出结果，得到一个多模型融合的人体行为识别器。基于 JHMDB 公开数据库的实验，结果证明了改进的双流 CNN 在人体行为识别任务上的有效性。

关键词：人体行为识别；深度学习；双流卷积神经网络；模型融合

An Improved Algorithm of Human Action Recognition Based on Two—Stream Convolutional Networks

Zhang Yijia, Mao Yaobin
(College of Automation, Nanjing University of Science & Technology, Nanjing 210094, China)

Abstract: In recent years, human action recognition has become a research hotspot in the field of computer vision. The research on convolutional neural network has made great breakthroughs in the field of image classification and recognition, but human action recognition is based on video, containing information in both spatial domain and temporal domain. Aiming at human action recognition based on video, an improved Two—Stream ConvNet architecture is proposed. For the spatial domain, the single RGB image is fed into the VGGNet _ 16 model. For the temporal domain, the superposition of multi optical flow images is fed into the Flow _ Net model. Finally, the Soft max outputs of the two models are merged with the linear weighting to realize human action recognition. The experiments based on the JHMDB public database prove the effectiveness of the improved two—stream ConvNet architecture on human action recognition.

Keywords: human action recognition; deep learning; two—stream convolution networks; model fusion

0 引言

人体行为识别的目的是分析并理解视频中的人体的动作和行为，与静态图像中二维空间的物体识别不同，行为识别主要研究如何感知目标对象在图像序列中的时空运动变化，将人体行为的表现形式从二维空间拓展到了三维时空。人体行为识别有着重要的理论意义且在很多领域有着重要的应用价值，如智能监控、视频检索和人机交互等[1]。

随着大规模数据集的涌现，传统算法已经很难满足如今大数据处理的需求，深度学习成为近几年国内外的研究热点。深度学习是机器学习领域的重点研究问题，它模拟人脑认知机制的多层次模型结构，通过组合低层特征形成更为抽象的高层特征来获得数据更有效的特征表示，相比于传统的人工提取特征更适合目标的检测和识别。

卷积神经网络是深度学习模型的典型代表，应用最为广泛，已经成为目前图像识别和语音分析等领域的一个应用热点。在人体行为识别方面，基于卷积神经网络的研究也有很多新进展。Ji 等人[2]在传统 CNN 基础上加入时间信息构成三维 CNN，将灰度、垂直和水平方向梯度、垂直和水平方向光流信息作为多通道输入，对于多个连续帧通过三维卷积操作实现视频数据在时间和空间维度的特征计算；Karpathy 等人[3]提出双分辨率的 CNN 模型，使用原始分辨率和低分辨率的视频帧分别作为输入，学习两个 CNN 模型，并在最后两个全连接层实现数据融合，以实现视频的最终特征描述用于后续识别；Karen 等人[4]提出双流 CNN 模型，将视频数据分成空间静态帧数据流和时域帧间动态数据流，分别将原始单帧 RGB 图像和多帧堆叠的光流图像分别作为两个 CNN 模型的输入进行特征提取，最后使用 SVM 分类器进行行为识别；Chéron 等人[5]提出使用根据人体姿势的关节点分割的单帧 RGB 图像和光流图像分别作为两个 CNN 模型的输入进行特征提取，并使用特征融合策略

收稿日期:2018-01-15； **修回日期**:2018-02-07。

作者简介:张怡佳(1993-)，女，江苏常熟人，硕士研究生，主要从事视频图像处理方向的研究。
茅耀斌(1971-)，男，江苏南京人，博士，副教授，主要从事视频图像处理方向的研究。

将视频数据转换为固定维度的特征向量，最后使用 SVM 分类器进行行为识别。

本文借鉴文献［4］中双流卷积神经网络模型中的“双流”概念，提出了一种基于改进双流卷积神经网络的人体行为识别模型，将 VGGNet _ 16 模型应用于双流卷积神经网络的空间流 CNN，替换原始的类 AlexNet 模型，从而加深网络结构；将 Flow _ Net 模型应用于双流卷积神经网络的时间流 CNN，替换原始的类 AlexNet 模型，使得模型更适用于提取光流图的特征，然后将空间流 CNN 模型和的时间流 CNN 模型的输出结果进行加权融合后作为双流 CNN 模型的输出结果，最终得到一个多模型融合的人体行为识别方法。

1　双流卷积神经网络

1.1　卷积神经网络

卷积神经网络[6]是一种特殊设计的深层模型，最早应用于图像识别领域。CNN 模型通过卷积和下采样操作自动学习图像特征，并把特征提取和分类输出合并为一个整体，从而获得更高的识别效率和更佳的性能表现。CNN 的核心思想是局部感受野、权值共享以及空间下采样，这使得网络的权值参数个数大幅减少，并获得了对图像位移、尺度、形变的不变性。典型的 CNN 网络结构如图 1 所示。

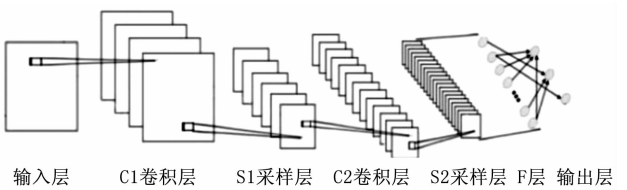

图 1　卷积神经网络结构示意图

这是一个简单的卷积神经网络，共有七层网络结构，其中基础层有卷积层、下采样层和全连接层，卷积层和下采样层是实现特征提取的关键，输出层采用 Softmax 分类器作类别判断。

1.1.1　卷积层

卷积层是通过多个不同的卷积核对上一层的输入做卷积运算得到多个输出，即多个特征图。卷积公式如式（1）所示：

$$x_j^l = f(\Sigma_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_i^l) \tag{1}$$

其中：x_j^l 表示第 l 层的第 j 个特征图，k_{ij}^l 表示第 $l-1$ 第 i 个特征图和第 l 层的第 j 个特征图对应的卷积核，b_i^l 表示第 l 层的第 i 个特征图对应的偏置，M_j 表示第 j 个特征图对应的输入特征图集合，f（·）表示一个激活函数，本文采用 ReLU 函数[7]。

1.1.2　下采样层

下采样层是对上一层的特征图进行采样操作，从而减小特征图的分辨率。采样操作是指对采样范围区域内所有像素点求平均值或最大值作为该区域采样后的值，从而实现卷积特征的降维并获得具有空间不变性的特征。本文采用最大值下采样操作，采样公式如式（2）所示：

$$y_{ij} = max_{0<h<H-1,\ 0<w+W-1}(x_{i\times H+h},\ y \times W + w) \tag{2}$$

其中：H，W 表示采样窗口的长和宽，x 表示二维输入向量，y 表示采样的输出值。

1.1.3　Softmax 分类器

深度学习网络常用的分类器包括多分类 SVM 以及 Softmax 分类器。本文选择使用 Softmax 作为特征提取后的多分类器。对于一个 k 分类任务，包含 m 个样本的训练集可表示为：

$$T = \{(x^{(1)},\ y^{(1)}),\ldots,\ (x^{(m)},\ y^{(m)})\} \tag{3}$$

其中：" $x^{(i)} \in R^{n+1}$ 表示一个 n 维向量的样本，$y^{(i)} \in \{1,\ 2,\ldots,\ k\}$ 是类别标签。对于输入样本 x，计算它属于每一个类别的概率：

$$P(y = j \mid x),\ (j = 1,\ldots,\ k) \tag{4}$$

Softmax 输出即为样本 $x^{(i)}$ 属于每个类别的所有概率值构成的一个 k 维的向量，计算函数如式（5）所示：

$$f(x^{(i)} \mid \theta) = \begin{bmatrix} P(y^{(i)} = 1 \mid x^{(i)},\ \theta) \\ P(y^{(i)} = 2 \mid x^{(i)},\ \theta) \\ \cdots \\ P(y^{(i)} = k \mid x^{(i)},\ \theta) \end{bmatrix} \tag{5}$$

其中：θ 是模型参数。

1.2　双流 CNN 网络结构

双流卷积神经网络的结构示意图如图 2 所示，该模型的核心在于空间流 CNN 和时间流 CNN 构成的“双流”结构，其中：空间流 CNN 以视频的单帧 RGB 图像作为输入，实现人体在空间域上表观信息的特征描述；而时间流 CNN 则是以多帧叠加后的光流图像作为输入，得到关于行为的运动特征表述，从而达到时间和空间互补的目的。针对给定的视频行为样本，首先分别通过时间流 CNN 和空间流 CNN 进行特征提取，最终将两个分支的分类结果进行加权融合，以得到关于视频中人体行为类别的最终决策结果。

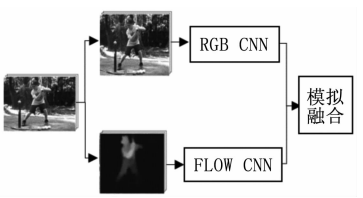

图 2　双流卷积神经网络模型结构示意图

原始双流卷积神经网络模型结构设计基本上和 AlexNet 模型是同一种思路，包括 5 层卷积层和 3 层全连接层，网络的输入图像尺寸被固定为 224×224。与 AlexNet 相比，原始双流 CNN 包含更多的卷积滤波器，第一层卷积层的卷积核尺寸缩小为 7×7，卷积步长减小为 2，其他层次的参数都与 AlexNet 相同。

随着对深度学习研究的深入，现在的网络结构发展呈现出层次结构更深，卷积核尺寸更小，滤波器数量更多，卷积操作步长更小的趋势，这些转变应用在物体检测任务上并获得了较好的效果。目前应用较广泛的深层次卷积神

经网络结构有 GoogleNet 和 VGGNet 和 ResNet 等。

本文选用 VGGNet－16 模型作为空间流 CNN 模型，VGGNet－16 是在数据库 ImageNet 上训练得到的具有 1000 个分类的模型，在 2014 年大规模视觉识别挑战赛（ILSVRC）中获得了第二名的成绩。VGGNet－16 模型继承了 AlexNet 模型的网络框架，采用了 16 层的深度网络，包含 13 个卷积层和 3 层全连接层，与 AlexNet 模型相比，VGGNet－16 模型使用了更深的网络，且所有卷积层都使用大小为 3×3 的卷积核，卷积步长也缩小到 1，能够模仿出更大的感受野，且减少了自由参数数目。VGGNet－16 模型结构如表 1 所示。

表 1 VGGNet－16 网络结构

层	卷积核	步长	下采样
卷积层	3×3×64	1	
卷积层	3×3×64	1	2×2
卷积层	3×3×128	1	
卷积层	3×3×128	1	2×2
卷积层	3×3×256	1	
卷积层	3×3×256	1	
卷积层	3×3×256	1	2×2
卷积层	3×3×512	1	
卷积层	3×3×512	1	
卷积层	3×3×512	1	2×2
卷积层	3×3×512	1	
卷积层	3×3×512	1	
卷积层	3×3×512	1	2×2
Fc－4096			
Fc－4096			
Fc－1000			
Softmax			

另外，时间流 CNN 用来提取光流信息，因此本文采用在光流图像上预训练的 Flow _ Net[8] 模型，Flow _ Net 模型是在包含 13320 个视频 101 类行为的 UCF101 数据库上训练光流图得到的模型，适合于用来进行光流图像的特征描述。Flow _ Net 模型的网络结构如表 2 所示。

表 2 Flow_Net 网络结构

层	卷积核	步长	下采样
层	卷积核	步长	下采样
卷积层	7×7×96	2	3×3
卷积层	5×5×384	2	3×3
卷积层	3×3×512	1	
卷积层	3×3×512	1	
卷积层	3×3×384	1	3×3
Fc－4096			
Fc－4096			
Fc－1000			
Softmax			

1.3 基于迁移学习的模型训练

为避免因训练样本数量不足出现深度网络学习过拟合的情况，本文采用迁移学习的方法。迁移学习[9-10]是指利用已学习到的知识解决不同但类似问题的方法，本文利用预训练好的模型初始化用于人体行为识别的双流卷积网络模型。

原始双流卷积神经网络只在空间流 CNN 上使用预训练，本文在空间流和时间流上分别使用 VGGNet－16 和 Flow _ Net 预训练模型初始化网络参数，并利用目标任务数据库对网络进行精调，得到目标任务网络模型。对于空间流 CNN，输入为 RGB 图像，而 VGGNet－16 模型由 ImageNet 数据库中的 RGB 图像训练得到，该数据库包含各种物体和生物，模型能够很好地获取图像低级、局部的特征，将其迁移到其他图片数据上的泛化性非常好，适合进行人体行为识别；而对于时间流 CNN，输入为光流图像，因此使用 UCF101 数据库上光流图像训练得到的 Flow _ Net 模型初始化时间流模型。另外，需要调小模型训练初始的学习率，并根据训练的迭代次数对学习率实时进行调整，随着迭代次数的增多，减小学习率。

1.4 模型融合

通过双流卷积神经网络获取视频帧的 RGB 特征以及光流特征后，需要将空间流 CNN 和时间流 CNN 的 Softmax 输出进行加权融合得到最终概率输出，选取概率最大的类别作为分类结果。

对于人体行为识别的分类任务，模型的输入是单帧图像，而样本是以单个行为视频为单位的，因此需要对视频中所有图像对应输出的概率矢量进行融合，得到某个视频单个模型的预测概率矢量，再将时间流以及空间流模型所得到的概率矢量以不同的权值相加，得到预测样本属于各个类别的概率向量 Vec：

$$Vec = \lambda \frac{\Sigma_n Vec_{\text{空间流}}}{n} + (1-\lambda) \frac{\Sigma_n Vec_{\text{空间流}}}{n} \tag{6}$$

其中：λ 是一个介于（0，1）的常量，n 是视频帧数。

2 实验结果与分析

2.1 实验平台与数据库

本文实验基于 Caffe 和 GPU，GPU 型号为 NVIDIA Titan X，该显卡的显存容量大小为 12G。

为验证模型的有效性，采用 JHMDB 人体行为数据库进行模型的性能测试，JHMDB 数据库是 HMDB51 的子数据集，包含 21 类行为，共有 928 个视频片段，视频帧分辨率为 320×240，其提供了训练集和测试集的划分，共有 3 个 splits。为满足网络训练需求，抑制过拟合，本文对图像样本做了一系列的数据扩充，通过对视频帧进行随机剪裁、随机旋转、水平翻转、对比度变化、亮度变化、加噪和模糊等处理，将图像样本扩充了 10 倍，同时增强了样本数据的多样性。

2.2 模型训练与结果分析

根据 JHMDB 数据库包含的行为类别数将 VGG－16 模

型和 Flow _ Net 模型的最后一个 Fc 层分类参数设置为 21；将 RGB 图像尺寸规范化到 224×224，光流图根据文献[10]计算得到，并将尺寸规范化到 227×227，每三帧光流图叠加作为一个输入样本，然后将单帧 RGB 原图和光流图像分别输入到 VGGNet－16 模型和 Flow _ Net 模型中，VGGNet－16 模型的初始学习率设为 0.001，每经过 10000 次迭代学习率降为原来的 10%，总共迭代 60000 次，Flow _ Net 模型的初始学习率设为 0.001，每经过 2000 次迭代学习率降为原来的 10%，总共迭代 10000 次，用测试集分别测试 VGG－16 模型和 Flow _ Net 模型。将两个模型得出的预测值进行融合，通过选取 5 种不同的权重融合，得出最终识别结果，表 3 为不同权重融合下得到的对 JHMDB 数据库中行为识别准确率的对比。

表 3　不同权重融合的效果比较

融合方法	识别率/%
单帧 RGB(λ＝1)	48.6
单帧光流(λ＝0)	52.3
1/3 空间流＋2/3 时间流(λ＝1/3)	60.14
1/2 空间流＋1/2 时间流(λ＝1/2)	58.25
2/3 空间流＋1/3 时间流(λ＝2/3)	57.43

从表 3 可以看出，时间流 CNN 比空间流 CNN 模型识别效果好，而经过模型融合得到的识别效果与不同模型预测结果的所占比重有关，总的来说，使用模型融合的方法要比单模型的分类效果好，且当 () 即空间流 CNN 模型和时间流 CNN 模型的输出以 1/3 和 2/3 的比重进行融合时，得到的最终分类结果效果最好，在 JHMDB 数据库 split1 上测试混淆矩阵如图 3 所示。

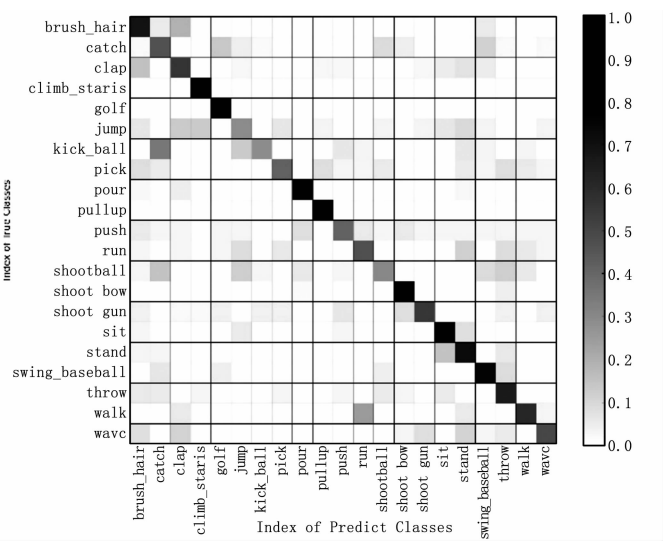

图 3　JHMDB－split1 双流 CNN 混淆矩阵

从图 3 可以看出，改进的双流 CNN 通过在新的数据库上进行微调，可以有效实现人体行为识别。其中，golf 的识别率最高，而 kick _ ball 的识别率最低，很容易被错分为 catch 或 jump。

本文提出的方法与其他人体行为识别方法的准确度进行对比，比较结果如表 4 所示。

表 4　与其他方法的效果比较

识别方法	识别率/%
原始双流 CNN[4]	59.4
P－CNN[5]	61.0
文献[8]方法	53.3
本文双流 CNN	60.14

从表 4 可以看出，本文提出的改进双流卷积神经网络相比于原始的双流卷积神经网络和文献[8]的方法在人体行为识别任务上的识别率略有提高。

3　结论

本文提出了一种改进的双流卷积神经网络模型，将 VGGNet _ 16 模型应用于空间流 CNN，替换原始的类 AlexNet 模型，从而加深网络结构；将 Flow _ Net 模型应用于时间流 CNN，替换原始的类 AlexNet 模型，使得模型更适用于提取光流图的特征，然后将空间流 CNN 和的时间流 CNN 的 Softmax 输出进行加权融合作为双流 CNN 模型的输出结果，最终实现人体行为识别。为了避免由于训练样本不足而出现模型过拟合现象，本文采用了训练样本集扩充和迁移学习的方法。最后，基于 JHMDB 数据库的实验得到改进的双流卷积神经网络模型的识别率达到 60.14%，证明了其在人体行为识别任务上的有效性。

参考文献：

[1] 单言虎，张　彰，黄凯奇. 人的视觉行为识别研究回顾、现状及展望[J]. 计算机研究与发展，2016，53(1)：93－112.

[2] 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.

[3] Karpathy A，Toderici G，Shetty S，et al. Large－scale video classification with convolutional neural networks [A]. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition [C]. 2014：1725－1732.

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

[5] Chéron G，Laptev I，Schmid C. P－CNN：Pose－based CNN features for action recognition [A]. Proceedings of the IEEE International Conference on Computer Vision [C]. 2015：3218－3226.

[6] 周飞燕，金林鹏，董　军. 卷积神经网络研究综述[J]. 计算机学报，2017，40 (6)：1229－1251.

(下转第 274 页)