

引用格式:毛志强,马翠红,崔金龙,等.基于双流卷积与双中心 loss 的行为识别研究[J].微电子学与计算机,2019,36(3):96-100.[MAO Zhi-qiang, MA Cui-hong, CUI Jin-long, et al. Research on action recognition based on two-stream convolution and double center loss [J]. Microelectronics & Computer, 2019, 36(3): 96-100.]

## 基于双流卷积与双中心 loss 的行为识别研究

毛志强<sup>1</sup>,马翠红<sup>1</sup>,崔金龙<sup>2</sup>,王毅<sup>1</sup>

(1 华北理工大学 电气工程学院,河北 唐山 063210;2 北京交通大学 海滨学院,河北 沧州 061100)

**摘 要:**针对行为视频中相似动作类内差异大、类间差异小,识别准确率不高的问题,提出了一种基于双流卷积网络与双中心 loss 的行为识别方法.该方法首先构建双流卷积网络结构,以 C3Dnet 模型作为双流结构的基础模型,分别提取多尺度 RGB 视频帧中的表观短时运动信息和堆叠光流图中的长时运动信息;然后将双流结构提取的深度信息经长短时记忆(LSTM)网络解析后进行特征融合;最后,利用基于双中心 loss 的 2C-softmax 目标函数,来最大化类间距离和最小化类内距离,从而实现相似动作的分类与识别.在数据集 KTH 上的实验结果表明,该方法能够准确识别相似动作,识别准确率可达 98.2%,具有很好的识别效果.

**关键词:** 双流卷积网络;中心 loss;长短时记(LSTM);光流图

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-7180(2019)03-0096-05

## Research on action recognition based on two-stream convolution and double center loss

MAO Zhi-qiang<sup>1</sup>, MA Cui-hong<sup>1</sup>, CUI Jin-long<sup>2</sup>, WANG Yi<sup>1</sup>

(1 College of Electrical Engineering, North China University of Science and Technology, Tangshan 063210, China;  
2 Beijing Jiaotong University Haibin College, Cangzhou 061100, China)

**Abstract:** Aiming at the problem of large difference in similar action classes, small difference between classes in action video and low recognition accuracy, a action recognition method based on two-stream convolution network and double-center loss is proposed. The method first constructs a two-stream convolutional network structure, and uses the C3Dnet model as the basic model of the two-stream structure to extract the apparent short-term motion information in the multi-scale RGB video frame and the long-term motion information in the stacked optical flow map respectively; Then, the depth information extracted by the two-stream structure is parsed by a long and short time memory (LSTM) network to perform feature fusion; Finally, the 2C-softmax objective function based on dual-center loss is used to maximize the distance between classes and minimize the distance within the class, so as to classify and identify similar actions. The experimental results on the data set KTH show that the method can accurately identify similar actions, and the recognition accuracy can reach 98.2%, which has a good recognition effect.

**Key words:** two-stream convolution network; center loss; long short-term memory (LSTM); optical flow map

1 引言

人体行为识别是计算机视觉领域非常活跃的研究课题,其广泛应用于视频安防监控、人机交互、环境控制检测等领域,涉及计算机视觉、模式识别、人工智能等多个学科<sup>[1-2]</sup>. 基于视觉的人体行为识别是对视频序列中人的行为进行分析和识别,是智能监控领域的一种关键技术.

传统的行为识别方法主要是基于手工特征的方法,然而由于人体行为种类较多,相似动作的类内类间差异,使得行为识别任务准确率难以提升. 近年来,随着深度学习方法在图像识别、自然语言处理等方面取得重大的研究进展,其在人体行为识别领域也得到了广泛的应用. 2014 年,Karpathy 等<sup>[3]</sup>首次利用深度卷积神经网络以连续的 RGB 视频帧为直接输入,进行行为识别;Simonyan 等<sup>[4]</sup>利用双流卷积网络分别提取堆叠光流图和 RGB 视频中的时空特征,虽然该模型某种程度上提取了时间维度信息,但双流结构均是 2D 卷积,识别效果并不是很好;2015 年,Cheron 等<sup>[5]</sup>利用 3D 卷积神经网络提取时间维度信息,是人体行为识别领域经典的模型;文献<sup>[6]</sup>,Du 等利用 C3Dnet 网络提取视频特征,并将其应用在行为识别、场景识别和视频相似度分析等领域,并取得了不错的效果. 但是以上模型,对于行为识别任务作用重大的时域动态信息利用并不充分,长短时记忆(Long Short-Term Memory,LSTM)网络是由递归神经网络(Recurrent Neural Network,RNN)演变而来<sup>[7]</sup>,能够递归学习输入序列中的长时动态信息. 然而,视频序列中人体行为识别最大的难点在于相似动作类间距离小、类内距离大,文献<sup>[8]</sup>提出了一个中心 loss 的解决方案. 中心 loss 通过对每一个行为识别学习一个中心,并根据样本特征与类中心的距离进行惩罚,从而减小类内距离,使得学习到的特征更具表征能力和泛化能力.

通过上述分析,本文结合各个网络模型优点,构建了一种基于双流卷积网络与双中心 loss 的行为识别模型.

2 模型架构设计

长时运动信息对于基于视频的人体行为识别准确率的提升具有重要的意义,然而,提取时间维度上的特征信息一直是动作识别领域中的难点. 因此,本文为了充分提取视频序列中的深度信息,引入 3D 卷积网络、长短时记忆(LSTM)网络和双流结构模

型,利用 C3Dnet 构建双流卷积网络模型分别提取多尺度 RGB 视频帧中的表观短时运动信息和堆叠光流图中的长时运动信息,再经过 LSTM 解析之后进行特征融合形成双流结构融合特征. 同时,为了进一步针对相似动作类间距离大、类内距离小的问题,设计了一种基于双中心 loss 的 2C-softmax 目标函数,其总体网络架构如图 1 所示. 该网络架构主要包括:双流 C3Dnet 模型提取视频特征、LSTM 解析视频特征、特征融合和 2C-softmax 目标函数.

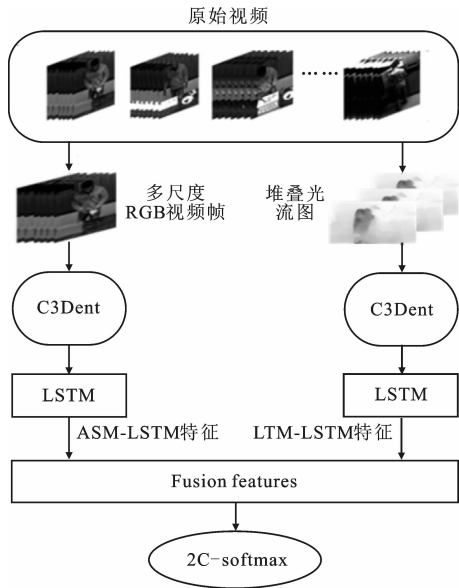


图 1 总体架构设计

2.1 双流 C3Dnet 提取视频特征

在原始的双流卷积网络模型及其改进模型中,一般均是采用二维卷积神经网络分别提取时间域运动信息和空间域表观信息,虽然在一定程度上提高了识别准确率,但是提取的视频序列特征的表征能力并不是很强,使得难以识别视频中人体相似动作. 文献<sup>[6]</sup>,将二维卷积扩展到三维,利用了时间维度信息,提出 C3Dnet 模型提取视频特征,并将其应用在行为识别、场景识别和视频相似度分析等领域. 通过 3D 卷积神经网络能够同时提取视频序列中的时空特征,使得表征能力增强. 虽然取得了一定效果,但是由于采用的是单流 C3Dnet 网络,使得提取的特征有限. 因此,本文在文献<sup>[4]</sup>的基础上,构建了双流 C3Dnet 模型分别提取多尺度 RGB 视频帧中的表观短时运动特征和堆叠光流图中的长时运动特征.

C3Dnet 模型包括 8 次卷积操作,每层卷积核大小均为  $3 \times 3 \times 3$ ,步长为 1;5 次池化操作,除第一层池化核大小为  $1 \times 2 \times 2$  外,其余池化核大小均为  $2 \times 2 \times 2$ ;2 个全连接层,每层输出响应均为 4 096 维,

其具体网络结构如图 2 所示。

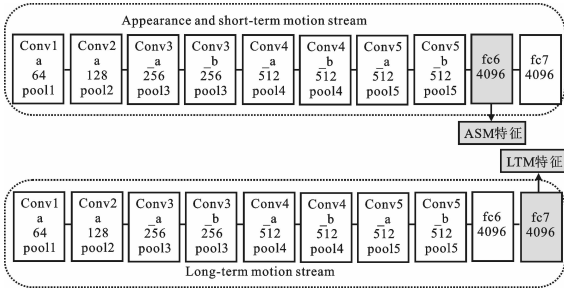


图 2 双流卷积网络结构

## 2.2 LSTM 网络结构

长短时记忆 (LSTM) 网络是由递归神经网络 (RNN) 演变而来, 对复杂的时间维度信息更加敏感, 能够有效地解决 RNN 在训练过程中出现的梯度爆炸或梯度弥散的问题, 从而有利于学习到长时动态信息. 其 LSTM 网络结构单元如图 3 所示。

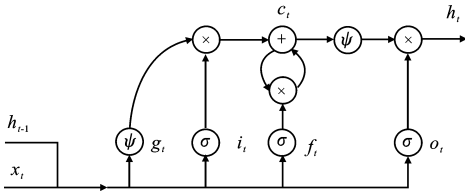


图 3 LSTM 结构单元

该结构单元在隐层中加入了先验知识——输入门、遗忘门、输出门和输入调制门, 通过以上门结构能够将各层间信号和某一时刻的输入信号处理的更加透明. 如式(1)所示:

$$\begin{aligned} i_t &= s(U_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= s(U_{xf}x_t + W_{hf}h_{t-1} + b_f) \\ o_t &= s(U_{xo}x_t + W_{ho}h_{t-1} + b_o) \end{aligned} \quad (1)$$

$g_t = y(U_{gx}x_t + W_{gh}h_{t-1} + b_g)$   
式中,  $U$  代表输入权重;  $W$  代表  $t-1$  时刻隐层递归权重;  $b$  为偏置项. 其中 LSTM 门结构由非线性激活函数控制, 设  $s(x) = (1 + e^{-x})^{-1}$  表示 sigmoid 函数,  $y(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  表示双曲正切函数.  $i_t$  决定保留哪种信息,  $f_t$  决定去除哪种信息,  $o_t$  确定记忆单元  $c_t$  哪部分将输出,  $g_t$  是由双曲正切函数创建的一个新的候选值向量。

记忆单元  $c_t$  是 LSTM 的核心部分, 其作用是选择有用信息去除无用信息, 如式(2) 所示:

$$c_t = f_t \times c_{t-1} + i_t \times g_t \quad (2)$$

式中,  $c_t$  包括两部分, 一部分是由上一时刻记忆单元  $c_{t-1}$  和遗忘门  $f_t$  相乘而得, 另一部分由输入门  $i_t$  和

输入调制门  $g_t$  相乘而得。

LSTM 的输出  $h_t$  是由输出门  $o_t$  控制是否激活记忆单元  $c_t$ , 如式(3) 所示。

$$h_t = o_t \times y(c_t) \quad (3)$$

视频行为识别的目的是从视频序列中识别出人体动作的时空视觉模式, 然而时序是复杂多变的. 通过双流 C3Dnet 结构提取到的特征经过 LSTM 网络解析后, 在时间域上会有更强的表征能力, 然后将其进行特征融合作为 2C-softmax 目标函数的输入, 从而完成人体行为的识别。

## 3 基于双中心 loss 优化网络

网络反向传播过程中是通过计算损失函数 (loss) 完成的, 在大多数的模型中一般是通过利用 softmax 的损失函数, 传统的 softmax-loss 如式(4) 所示:

$$L_j = - \sum_{i=1}^m \log \frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T x_i + b_j}} \quad (4)$$

式中,  $x_i$  表示第  $i$  个特征向量;  $y_i$  表示类别标签;  $n$  为类别数;  $m$  表示小批量大小;  $W$  表示权重;  $b$  为偏置项。

文献[8]中 Wen 等人设计了中心 softmax-loss 函数用于人脸识别任务, 将特征空间中的每一个类别都保持一个类中心  $C$ , 如图 4(a) 所示. 具体而言, 中心 loss 同时学习每个类别的深层特征的中心  $C$ , 并惩罚深层特征与其相应的类别中心之间的距离, 从而能够减小类内距离并扩大类间距离. 文中 loss 函数包含 softmax loss 和中心 loss 两部分, 其中心 softmax-loss 的计算公式如式(5) 所示:

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|^2 \quad (5)$$

$$L = L_j + \lambda L_c$$

式中,  $x_i$  表示第  $i$  样本特征向量;  $c_{y_i}$  表示该样本所属类别的特征值中心;  $L_c$  表示中心 loss 计算公式;  $L_j$  为 softmax 函数的 loss;  $\lambda$  为两者所占比重。

由于本文是利用双流网络结构提取的两类运动特征, 因此, 本文在中心 loss 的基础上设计了双中心 loss, 如图 4(b). 双中心 loss 分别维护表观短时运动流特征中心  $C_{AS}$  和长时运动流特征中心  $C_{LT}$ , 两者按一定权重系数  $W_{AS}$  和  $W_{LT}$  融合形成质心  $C_i$ . 本文采用线性加权方式确定权重系数  $W_{AS}$  和  $W_{LT}$ , 使质心  $C_i$  在  $C_{AS}$  和  $C_{LT}$  的连线之间, 从而能够保证质心  $C_i$  同时离两者之间距离最近. 如式(6) 所示:

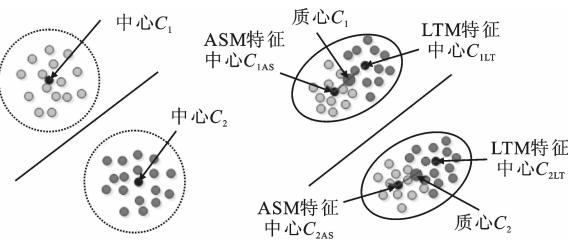


图 4 中心 loss 的特征值划分

$$L = L_j + W_{AS}L_{C_{AS}} + W_{LT}L_{C_{LT}} \tag{6}$$
式中, $W_{AS}$ 和 $W_{LT}$ 为双中心 loss 的权值系数; $L_{C_{AS}}$ 表示 ASM 特征中心 loss; $L_{C_{LT}}$ 表示 LTM 特征中心 loss.

为了防止目标函数过拟合,可以在 2C-softmax 的 loss 中加入正则项. 在网络结构中,融合特征 (Fusion features) 单元对整个行为识别过程具有巨大的影响,所以能够加入融合特征单元权值的二范数作为正则项,如式(7)所示:

$$L_F = \frac{1}{2} \sum_{i=1}^m \|W_{F_i}\|_2^2 \tag{7}$$

式中, $m$ 为小批量的大小; $W_{F_i}$ 为第*i*个特征样本的权值; $F$ 表示融合特征单元个数. 那么式(6)可以改写为:

$$\bar{L} = L_j + W_{AS}L_{C_{AS}} + W_{LT}L_{C_{LT}} + \alpha L_F \tag{8}$$

式中, $\alpha$ 为正则项系数.

4 实验结果与分析

4.1 实验环境搭建

本文实验环境选择深度学习框架 caffe 平台设计实现,网络训练采用小批量随机梯度下降法,动量为 0.9,权值衰减率为 0.000 5. 双流卷积网络结构采用 C3Dnet 网络作为基础模型,以 16 帧为一组的堆叠 RGB 视频帧作为表观短时运动流,初始学习率为 0.005,以 10 对(flow\_x 和 flow\_y)光流图堆叠而成的光流组作为长时运动流输入. 针对堆叠光流图的计算采用 OpenCV 视觉库中的稠密光流帧提取方法,分别提取水平方向和垂直方向的光流序列. 然后,将双流结构提取的 AS-fc6 层特征与 LT-fc7 层特征通过 LSTM 网络解析,得到更具表征能力的深度特征,其学习率为 0.001,梯度阈值为 15,批次大小为 100,训练周期为 150,LSTM 单元隐结点数为 18. 对于 2C-softmax 的误差正则化系数  $\alpha=0.000\ 1$ , $W_{AS}=0.000\ 07$ , $W_{LT}=0.000\ 03$ .

本次实验数据集采用 KTH 数据集,该数据集

包括了 4 种场景下 25 个不同行人的 6 中行为视频: 正常行走 (Walk)、慢跑 (Jog)、跑 (Run)、挥拳 (Box)、双手挥手 (Wave)、鼓掌 (Clap).

4.2 实验结果

实验过程中,为了增加识别准确率可信度,本文将 KTH 数据集随机划分成 3 组,取其 3 组测试平均准确率作为评估模型效果的指标.

首先,本文利用 LSTM 网络对双流网络结构的输出特征分别进行解析,并对比了网络层输出特征的不同方式的行为识别准确率,如表 1 所示.

表 1 不同网络层输出特征识别准确率的比较 %

| Layer of output feature | KTH  |
|-------------------------|------|
| AS-Conv5 & L-Conv5      | 88.2 |
| AS-fc6 & L-fc6          | 95.3 |
| AS-fc7 & L-fc7          | 96.2 |
| AS-fc7 & L-fc6          | 92.7 |
| AS-fc6 & L-fc7          | 98.2 |

从表 1 中可以发现,当双流 C3Dnet 结构的输出特征的层为卷积层时,行为识别效果并不是很好.

接下来,为了验证本文算法提出的双中心 loss 对网络模型的作用,在 KTH 数据集上对比了不同的 loss 设计方案对网络的影响. 实验结果表明,单纯地采用双中心 loss 对行为识别效果并不大,而加入融合单元的正则项后使得识别准确率得到很明显的提升. 如表 2 所示.

表 2 不同 Loss 方案对网络的影响

| Loss                                                     | 准确率/% |
|----------------------------------------------------------|-------|
| $L_j + \alpha L_c$                                       | 92.7  |
| $L_j + W_{AS}L_{C_{AS}} + W_{LT}L_{C_{LT}}$              | 93.1  |
| $L_j + W_{AS}L_{C_{AS}} + W_{LT}L_{C_{LT}} + \alpha L_F$ | 98.2  |

此外,通过训练数据集对网络模型进行训练后,再通过测试数据集对本文模型进行测试,可以得到各种行为识别的混淆矩阵,如表 3 所示.

表 3 各种行为的混淆矩阵

| 行为   | Walk        | Jog         | Run         | Box         | Wave        | Clap        |
|------|-------------|-------------|-------------|-------------|-------------|-------------|
| Walk | <b>0.96</b> | 0.03        | 0           | 0.009       | 0           | 0           |
| Jog  | 0.02        | <b>0.97</b> | 0.01        | 0           | 0           | 0           |
| Run  | 0.01        | 0.01        | <b>0.99</b> | 0           | 0           | 0           |
| Box  | 0           | 0           | 0           | <b>0.98</b> | 0           | 0.01        |
| Wave | 0           | 0           | 0           | 0           | <b>1.00</b> | 0           |
| Clap | 0           | 0           | 0           | 0.01        | 0           | <b>0.99</b> |

从表 3 中可以看出,本文提出的算法能够识别出绝大多数的行为动作.对于 Jog 与 Walk、Run 是最容易混淆的,属于相似动作,然而识别的误识别率只有 0.01~0.03.因此,通过混淆矩阵可以看出,本文算法针对相似动作行为具有很好的识别效果.

最后,将本文算法与现今主流的行为识别算法<sup>[4,6,9]</sup>在 KTH 数据集上进行比较,如表 4 所示.

| 表 4 不同算法在 KTH 上的比较结果 % |       |       |       |                 |      |
|------------------------|-------|-------|-------|-----------------|------|
| 算法                     | 文献[4] | 文献[6] | 文献[9] | Ours(无 LSTM 解析) | Ours |
| Walk                   | 93    | 94    | 93    | 93              | 96   |
| Jog                    | 87    | 91    | 82    | 91              | 97   |
| Run                    | 86    | 89    | 79    | 97              | 99   |
| Box                    | 100   | 99    | 95    | 99              | 98   |
| Wave                   | 97    | 98    | 99    | 100             | 100  |
| Clap                   | 96    | 92    | 93    | 92              | 99   |
| 识别率                    | 93.2  | 93.8  | 90.1  | 95.3            | 98.2 |

从表 4 中可以发现,本文提出的基于双流卷积网络与双中心 loss 的行为识别效果优于其他算法,而且通过加入 LSTM 解析双流结构特征后,能够提高识别准确率.

### 5 结束语

本文提出的基于双流卷积网络和双中心 loss 的行为识别方法,从相似动作行为类内类间距离差异角度出发,通过双流 C3Dnet 结构分别提取多尺度 RGB 视频帧中的表观短时运动特征和堆叠光流图中的长时运动光流特征;然后采用 LSTM 网络解析双流结构 AS-fc6 层和 L-fc7 层的输出特征,并进行特征融合;最后,利用基于双中心 loss 设计的 2C-softmax 目标函数对网络模型进行优化,实现行为识别任务.在 KTH 数据集上的实验结果表明,本文算法的识别效果与其他算法相比更具优势,而且针对某些相似动作也具有很高的识别准确率.

### 参考文献:

[1] 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.

[2] HERATH S, HARANDI M, PORIKLI F. Going deeper into action recognition: a survey [J]. Image &

Vision Computing, 2017, 60(4):4-21.

[3] KARPATY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio, IEEE Computer Society, 2014:1725-1732.

[4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Computer Vision and Pattern Recognition, 2014, 1(4):568-576.

[5] CHÉRON G, LAPTEV I, SCHMID C. P-CNN: pose-based cnn features for action recognition[C]// IEEE International Conference on Computer Vision. Santiago, Chile, IEEE Computer Society, 2015:3218-3226.

[6] DU T, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// IEEE International Conference on Computer Vision. Rome, Italy, IEEE, 2016:4489-4497.

[7] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4):677-691.

[8] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition [C]// Computer Vision, ECCV 2016. Amsterdam, The Netherlands, Springer International Publishing, 2016:499-515.

[9] KARPATY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA, IEEE Computer Society, 2014:1725-1732.

### 作者简介:

毛志强 男,(1991-),硕士研究生.研究方向为计算机视觉、目标检测与人体行为分析.

马翠红(通讯作者) 女,(1960-),教授.研究方向为复杂工业系统的建模与控制、图像识别与视频分析.

E-mail:864404484@qq.com.

崔金龙 男,(1989-),硕士,助教.研究方向为钢成分测量.

王毅 男,(1994-),硕士研究生.研究方向为计算机视觉、目标检测与视频分析.