DOI:10.19651/j.cnki.emt.2209490

基于稠密连接时空双流网络的行为识别方法研究*

程换新[1] 孙胜意[1] 骆晓玲[2] 王 雪[1]
(1.青岛科技大学自动化与电子工程学院 青岛 266061; 2.青岛科技大学机电工程学院 青岛 266061)

摘 要:针对视频中复杂人体动作识别精度低、效率差的问题,提出了一种时空特征提取的稠密连接网络模型。首先利用两个稠密连接网络进行时空特征的提取;其次构建时空网络间的稠密连接,将时间网络中提取到的特征信息逐层输入到空间流网络中,提高两个流的时空交互性;然后使用 LSTM 网络分别对双流网络特征进行处理得到两个流的预测结果;最后融合双流网络的预测结果,从而实现视频中复杂行为的识别。在 UCF101 和 HMDB51 两个基准数据集上进行对比实验,得到 94.69%和 68.87%的准确率,优于其他算法。实验证明,本文模型可增加时空网络之间的交互性,有利于对复杂人体动作的识别。

关键词:时空双流模型;稠密网络;人体动作识别;LSTM
中图分类号:TP391.9 **文献标识码**:A **国家标准学科分类代码**:520.60

Research on behavior recognition method based on densely connected spatiotemporal two-stream network

Cheng Huanxin[1] Sun Shengyi[1] Luo Xiaoling[2] Wang Xue[1]
(1. College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China;
2. College of Electrical Mechanical Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Aiming at the problems of low accuracy and low efficiency of complex human motion recognition in video, a dense connection network model for spatio-temporal feature extraction is proposed. Firstly, two dense connected networks are used to extract spatiotemporal features; Secondly, the dense connection between spatiotemporal networks is constructed, and the feature information extracted from the spatiotemporal network is input into the spatial flow network layer by layer to improve the spatiotemporal interaction between the two flows; Then the LSTM network is used to process the characteristics of the two stream network respectively, and the prediction results of the two streams are obtained; Finally, the prediction results of dual stream network are fused to realize the recognition of complex behaviors in video. The comparative experiments on ucf101 and hmdb51 benchmark data sets show that the accuracy rates of 94.69% and 68.87% are better than other algorithms. Experiments show that this model can increase the interaction between spatiotemporal networks and is conducive to the recognition of complex human actions.

Keywords: spatiotemporal two stream model; human action recognition; dense network; LSTM

0 引 言

近些年,随着人工智能的发展,对计算机视觉领域的研究越来越引起人们的重视,对人体行为识别的研究也是其中的热点话题。行为识别的主要目的是针对给定的一段视频,识别视频中的人类行为并对其进行分类。目前的视频行为识别算法分为手工特征提取方法[1-3]和基于神经网络的方法[4-8]。

行为识别方法研究多年,传统算法主要是通过传统机器学习方法对视频特征进行手工特征提取。最经典的网络是 Wang 等[9]于 2013 年提出的密集轨迹(dense trajectory, DT)算法,该方法可有目的性的提取特征信息。但其在复杂环境下对特征的提取效果很差,识别效率低,无法满足实时识别的要求。Wang 等[10]于 2014 年将 DT 算法进行改进,研究出提升的密集轨迹算法(improved dense trajectories,iDT)算法。iDT 算法在手工特征提取方法中效果最好,其加入了消除背景光流的方法,可以更好的对人体运动特征进行提取,减少了复杂环境下的外界干扰,但运

收稿日期:2022-04-02
* 基金项目:国家海洋局重大专项(国海科字[2016]494 号)资助

算速度慢、识别效率低的问题仍然无法解决。因此需要用到基于深度学习的方法进行行为识别。

近几年,深度学习被广泛用于各个行业。基于深度学习的方法主要通过神经网络模拟人类各种行为,可以模仿人的视觉原理保存图像特征,因此被许多科研人员应用到行为识别研究中。Simonyan 等[11]于 2014 年研究出一种新的模型,在行为识别方面有较大影响。该模型为时空双流卷积网络(Two-Stream)。网络框架是由空间和时间两个维度的网络组成,一部分处理空间图像,另一部分处理光流图像,最终联合训练,对行为结果进行识别分类,但该算法采用直接平均的方法对分支网络结果进行融合,融合结果较差。针对时空双流网络行为识别效果差的问题,Feichtenhofer 等[12]在时空双流网络的基础上进行改进,使用卷积神经网络(convolutional neural networks,CNN)进行双流网络结果的融合,提高了识别效果。Wang 等[13]为了提高双流网络长时间建模的能力,研究出新的时间分段网络(time-sensitive network,TSN)网络。该网络将一段视频按相等间隔分为多个部分,从每个部分中随机选取一个视频段进行时空双流卷积训练,对多个视频段上提取的特征进行融合,进行视频级预测,从而解决时空双流网络无法对长视频建模的问题。Zhou 等[14]对 TSN 网络进行改进,通过在时序上对视频帧进行推理,可以识别更多复杂动作。Yue 等[15]在双流网络融合阶段加入长短期记忆网络(long short-term memory,LSTM),提高了对复杂运动的识别能力,但由于该网络特征提取能力不足导致识别效率降低。

将双流网络与其他模型结合的改进算法固然很好,但除了最后的融合层,大多数基于双流网络的行为识别算法的两个流往往是完全独立没有任何联系的。在这种情况下缺乏时间流和空间流的交互,会影响模型的性能。因此需要在两个流之间建立稠密连接,增加两个流的交互性,能有效提高模型性能。通过上述分析,本文构建稠密连接时空双流网络来提高模型性能。

1 网络结构设计

本文提出了基于稠密连接的时空双流网络模型,该模型整体结构如图 1 所示。

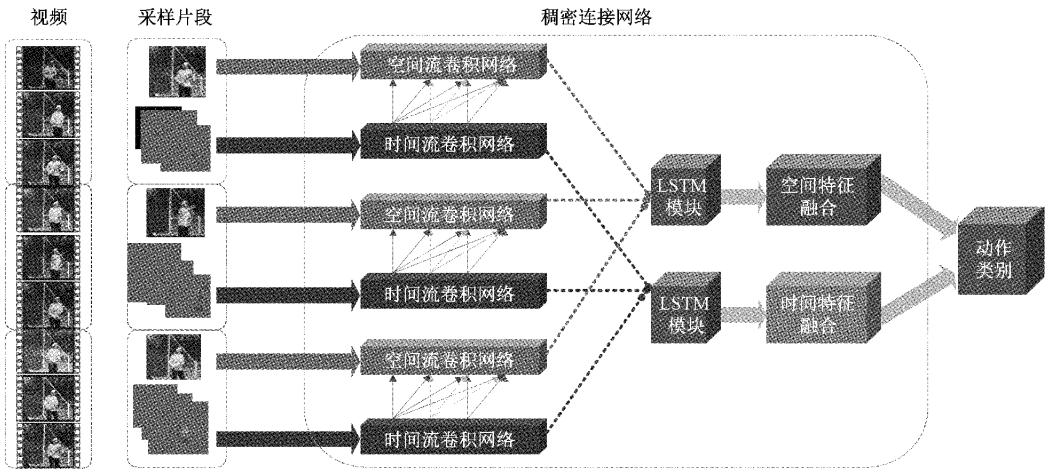

图 1 稠密连接时空双流卷积网络模型

模型由 4 个部分组成:时空特征提取部分、双流网络之间稠密连接、基于 LSTM 的时空预测模块、时空融合模块。首先采用两个稠密连接网络(DenseNet)对时空特征进行提取;其次采用稠密连接方法把时间流和空间流网络联系起来,将时间流提取到的特征信息逐层输入到空间流网络中,能够在保证两个流信息特异性的同时,提高两个流的时空交互性;然后使用 LSTM 分别对双流网络提取到的特征进行预测;最后对两个流的结果进行融合,从而实现视频动作识别。

1.1 稠密连接卷积网络

本文所用的稠密连接卷积网络模型是由 Huang 等[16]于 2017 年提出的 DenseNet121,模型结构如图 2 所示。该模型由多个稠密块组成,内部每层间均是直接连接,从而增加每层输入的多样性。每个稠密块之间由过渡层连接,过渡层是由卷积和池化层组成的,可以通过下采样变换特征图的大小。

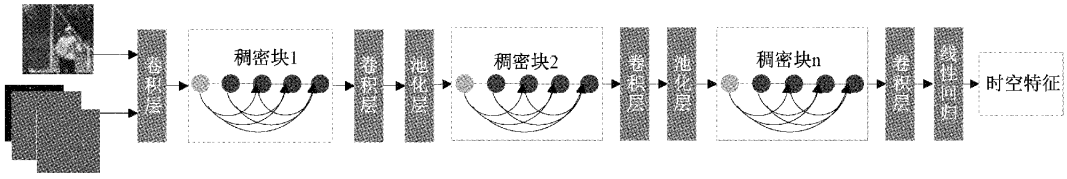

图 2 DenseNet121 网络结构

本文时空双流网络均采用 DenseNet121 网络模型，但两个流的输入信息不同。空间网络主要输入 RGB 图像，通过对静态图像中的特征信息进行提取，可以对视频中的动作进行分类识别。而时间网络则以光流图作为输入，光流图可看做位移矢量场，矢量场为双通道特征，包含水平和垂直两个部分，从而更好地描述运动轨迹，更有利于提取人体动作信息。该网络可以利用所用先前层的特征图作为当前层的输入，可以有效缓解梯度消失问题。同时对于特征传输和特征重用方面也优于其他网络模型。

1.2　双流网络间的稠密连接

双流网络间的稠密连接结构如图 3 所示，空间流网络的输入是时间流网络的所有前在块的特征图和空间流的前一块特征图。首先不同层次的稠密块对各个视频段的输出特征进行提取；然后通过乘法门调整时空特征用来过滤信息，从而得到更有价值的补偿信息，同时强制两个流在前馈通道和反馈通道中相互作用；最后通过逐元素加法将有价值的补偿信息输入到下一个稠密块实现双流网络之间的信息交流。

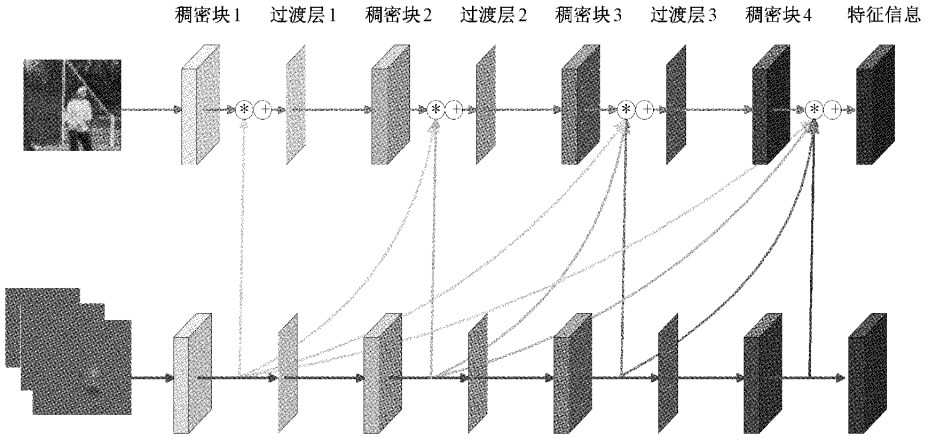

图 3　双流网络间的稠密连接网络结构

其中符号“＋”和符号“＊”分别表示逐元素加法和逐元素乘法，具体方法如式(1)～(3)所示。

$$\boldsymbol{X}_R^{i+1}=h(\boldsymbol{X}_R^i)+G(h(\boldsymbol{X}_R^i),\boldsymbol{X}_T^i)) \tag{1}$$

$$G(h(\boldsymbol{X}_R^i),\boldsymbol{X}_T^i))=h(\boldsymbol{X}_R^i)*[H^i(X_T)] \tag{2}$$

$$[H^i(X_T)]=[\boldsymbol{H}_1^i(\boldsymbol{X}_T^1),\cdots,\boldsymbol{H}_i^i(\boldsymbol{X}_T^i)] \tag{3}$$

其中，$\boldsymbol{X}_R^{i+1}$ 和 $\boldsymbol{X}_F^i$ 表示第 i 个稠密块和过渡块的输入，$h(\boldsymbol{X}_i^R)$表示将空间流第 i 个块的输入传递到相应卷积层中的第 $i+1$ 个块的原始函数，G 表示乘法门，可以同来调整空间特征和时间特征的融合信息，$[H^i(X_T)]$表示时间流所有前在稠密块特征图的连接，$\boldsymbol{H}_j^i$ 表示权重矩阵，可以将 $\boldsymbol{X}_T^i$ 转换成和 $\boldsymbol{X}_R^i$ 相同的大小。

1.3　基于 LSTM 的时空特征预测

由于动作的不同阶段的特征对动作识别的贡献不同，在特征信息经过稠密网络提取时，不同采样片段提取的信息价值也不相同。对融合结果的贡献率大的特征信息为高价值信息，应受到更多关注，因此本文采用长短期记忆网络(LSTM)对稠密网络提取的时空特征进行处理，以时空特征信息作为输入，输出当前时刻的视频级预测结果。LSTM 网络结构如图 4 所示。

该模型中加入了输入门、遗忘门和输出门，通过门结构能够对某一时刻输入的信号进行处理。每个门的方法如式(4)～(9)所示。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{4}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{5}$$

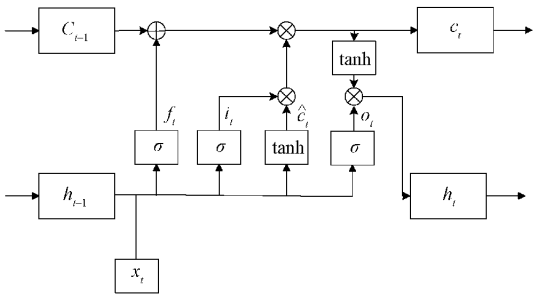

图 4　LSTM 网络结构

$$\tilde{C}_t=\tanh(W_C\cdot[h_{t-1},x_t]+b_C) \tag{6}$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t \tag{7}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{8}$$

$$h_t=o_t*\tanh(C_t) \tag{9}$$

如公式所示，$\boldsymbol{W}$ 表示对应的权重向量，b 是偏置项。i_t 是输入门，可以决定所保留的信息。f_t 是遗忘门，通过记忆单元可以去除无用信息。C_t 表示记忆单元，结合候选值信息 $\tilde{C}_t$ 对细胞状态进行更新。o_t 是输出门，与记忆单元相结合输出结果 h_t。

2　实　　验

2.1　实验设计

本实验在 Windows10 操作系统上进行，CPU 为 AMD Ryzen 5 3600 6-Core Processor 3.59 GHz，GPU 为 GeForce

GTX 2060,深度学习框架选用 PyTorch 平台。

训练方法采用随机梯度下降法(stochastic gradient descent,SGD)。模型使用预训练的 DenseNet121 模型进行特征提取,首先将等间隔采样的视频帧输入到空间卷积网络中,尺寸设定为 224 * 224;使用 OpenCV 模块对光流进行提取,作为时间卷积神经网络的输入,尺寸设定为 224 * 224 * 2L。对双流网络的训练参数进行设定,批尺寸设定为 32,动量因子设定为 0.85。迭代次数设定为 60 000 代,学习率由 0.01 开始每迭代 10 000 次递减 1/10,直至停止训练。

实验数据集为基准数据集 UCF101[17] 和 HMDB51[18]。网络模型分别在这两个数据集上进行实验,以验证网络泛化性。UCF101 数据集种类丰富,视频数量为 13 320 个,而 HMDB51 数据集是从各种来源的真实视频中收集的,视频共计 6 766 个。HMDB51 数据集是从各种视频网站获取的,因此识别动作较复杂,相对于 UCF101 数据集更具有识别难度更大。在实验中训练集和测试集按照 3∶1 的比例分割。

2.2 实验结果分析

本实验所提出的稠密连接时空双流网络模型主要目的是为了提高动作识别的精度,是通过对 TSN 网络改进提出的,因此将本文方法与 TSN 网络及不同改进方法在 UCF101 和 HMDB51 数据集上进行对比实验,所得结果如表 1 所示。

表 1 不同改进方法的对比实验结果 %

改进方法	UCF101	HMDB51
TSN	92.64	65.74
TSN+LSTM	93.36	66.85
TSN+稠密连接+LSTM	**94.69**	**68.87**

由表 1 中可以看出,在特征提取之后加入 LSTM 结构,动作识别的准确率有所提升,在 UCF101 数据集上提升了 0.7%,对于视频复杂度更高的 HMDB51 数据集提升了 1.1%,这表明在特征提取后加入 LSTM 可以提升动作识别的准确率。采用稠密连接对于准确率的提升效果更好,在 UCF101 和 HMDB51 数据集上分别提升 2.0%和 3.1%,这说明加入稠密连接增强双流网络的交互性可以提高动作识别的准确率。

最后,将本文算法与当前主流算法进行对比实验,实验结果如表 2 所示。由表 2 可以看出,本文提出的方法优于目前主流方法,并且对于复杂度较高的视频识别效果更高。分别在 UCF101 数据集和 HMDB51 数据集上将动作识别准确率提高了 0.8%和 1.9%。因此本文算法可以有效提高动作识别的精度,具有一定的优越性。

表 2 本文模型与其他模型的对比实验结果 %

主流方法	UCF101	HMDB51
iDT[10]	85.93	57.2
Two-Stream[11]	88.02	59.4
TSN[13]	92.64	65.74
C3D[19]	93.43	66.81
STDDCN[20]	93.81	66.93
本文方法	**94.69**	**68.87**

3 结 论

本文提出了一种基于稠密连接的时空双流卷积网络的视频行为识别算法。首先通过稠密连接网络对时空特征进行提取;其次建立双流网络间的稠密连接结构,提高双流网络间的联系;然后使用 LSTM 网络对双流网络提取的特征进行处理得到两个流的预测结果;最后进行时空网络结果融合,实现人体动作识别。在基准数据集上的实验表明:使用稠密连接网络增加双流网络间的交互性能够更好的提高双流网络间的信息传递,并且使用 LSTM 对特征进行处理有更好的识别效果。本文算法相交于其他主流算法在 UCF101 和 HMDB51 数据集上准确率分别提升了 0.7%和 1.8%。表明本文方法的识别效果比其他主流方法效果更好,并且在复杂度较高的视频识别上更具有优势。再下来的工作中还会进一步改进网络,提高网络在光照效果较差、能见度较低的环境中的动作识别效果。

参考文献

[1] DUTA I C, IONESCU B, AIZAWA K, et al. Spatio-temporal vlad encoding for human action recognition in videos[C]. International Conference on Multimedia Modeling. Springer, Cham, 2017: 365-378.

[2] YANG X, TIAN Y L. Action recognition using super sparse coding vector with spatio-temporal awareness[C]. European Conference on Computer Vision. Springer, Cham, 2014: 727-741.

[3] PENG X, ZOU C, QIAO Y, et al. Action recognition with stacked fisher vectors[C]. European Conference on Computer Vision. Springer, Cham, 2014: 581-595.

[4] GUO Y, WANG X. Applying TS-DBN model into sports behavior recognition with deep learning approach[J]. The Journal of Supercomputing, 2021, 77(10): 12192-12208.

[5] CAO B, XIA H, LIU Z. A video abnormal behavior recognition algorithm based on deep learning[C]. 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference(IMCEC), IEEE, 2021, 4: 755-759.

[6] ATTO A M, BENOIT A, LAMBERT P. Timed-image based deep learning for action recognition in video sequences[J]. Pattern Recognition, 2020, 104, DOI:10.13140/RG.2.2.12648.11526/1.

[7] 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.

[8] MIRZA A, SIDDIQI I. Recognition of cursive video text using a deep learning framework[J]. IET Image Processing, 2020, 14(14): 3444-3455.

[9] WANG H, KLÄSER A, SCHMID C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International Journal of Computer Vision, 2013, 103(1): 60-79.

[10] WANG H, SCHMID C. Action recognition with improved trajectories[C]. Proceedings of the IEEE International Conference on Computer Vision, 2014: 3551-3558.

[11] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, DOI: 10.1002/14651858.CD001941.pub3.

[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] 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.

[14] ZHOU B, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2018: 803-818.

[15] YUE-HEI NG J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4694-4702.

[16] HUANG G, LIU Z, LAURENS V, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[17] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012, DOI: 10.48550/arXiv.1212.0402.

[18] 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.

[19] 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.

[20] HAO W, ZHANG Z. Spatiotemporal distilled dense-connectivity network for video action recognition[J]. Pattern Recognition, 2019, 92: 13-24.

作者简介

程换新，博士，教授，主要研究方向为控制科学与工程、人工智能、图像识别等。
E-mail: 2635510239@qq.com

孙胜意，研究生，主要研究方向为人工智能、图像识别等。
E-mail: 1932510636@qq.com

骆晓玲(通信作者)，博士，教授，主要研究方向为过程设备及控制的设计研究等。
E-mail: xiaolingluo@126.com

王雪，研究生，主要研究方向为人工智能、图像识别等。
E-mail: 2846643978@qq.com