计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):234−239 [doi: 10.15888/j.cnki.csa.006669]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 RGB-D 视频的多模态手势识别①

马正文[1], 蔡坚勇[1,2,3,4,5], 刘　磊[1], 欧阳乐峰[1], 李　楠[1]

[1](福建师范大学 光电与信息工程学院, 福州 350007)
[2](福建师范大学 医学光电科学与技术教育部重点实验室, 福州 350007)
[3](福建师范大学 福建省光子技术重点实验室, 福州 350007)
[4](福建师范大学 福建省光电传感应用工程技术研究中心, 福州 350007)
[5](福建师范大学 智能光电系统工程研究中心, 福州 350007)
通讯作者: 蔡坚勇, E-mail: cjy@fjnu.edu.cn

摘　要: 本文是对 SKIG RGB-D 多模态的孤立手势视频进行手势识别研究.首先将 RGB 和 Depth 两种单模态视频提取成图片的形式保存, 然后采样成长度为 32 帧的手势序列分别输入到本文提出的稠密连接的 3DCNN 组件学习短期的时空域特征, 然后将提取的时空域特征输入到卷积 GRU 网络进行长期的时空域特征学习, 最终对单模态训练好的网络进行多模态融合, 提升网络识别准确率.本文在 SKIG 数据集上取得了 99.07% 的识别准确率, 达到了极高的准确率, 证明了本文提出的网络模型的有效性.

关键词: 手势识别; 稠密连接的 3DCNN; 卷积 GRU; 时空域特征

引用格式: 马正文,蔡坚勇,刘磊,欧阳乐峰,李楠.基于 RGB-D 视频的多模态手势识别.计算机系统应用,2018,27(12):234–239. http://www.c-s-a.org.cn/1003-3254/6669.html

Multimodal Gesture Recognition Based on RGB-D Video

MA Zheng-Wen[1], CAI Jian-Yong[1,2,3,4,5], LIU Lei[1], OUYANG Le-Feng[1], LI Nan[1]

[1](College of Photonic and Electronic Engineering, Fujian Normal University, Fuzhou 350007, China)
[2](Key Laboratory of Optoelectronic Science and Technology for Medicine of Ministry of Education, Fujian Normal University, Fuzhou 350007, China)
[3](Fujian Provincial Key Laboratory of Photonics Technology, Fujian Normal University, Fuzhou 350007, China)
[4](Fujian Provincial Engineering Technology Research Center of Photoelectric Sensing Application, Fujian Normal University, Fuzhou 350007, China)
[5](Intelligent Optoelectronic Systems Research Center, Fujian Normal University, Fuzhou 350007, China)

Abstract: In this study, the gesture recognition based on SKIG RGB-D multimodal isolated gesture video is studied. The RGB and depth videos are extracted into the form of images. Then the sampled 32 frames from images are input to the densely connected 3DCNN component to learn short-term spatiotemporal features, after that the features input to the convolutional GRU to learn long-term spatiotemporal features. Finally, the trained networks for single modal are used to multimodal fusion to improve the recognition accuracy. 99.07% recognition accuracy is obtained on the SKIG dataset, which achieves high accuracy and proves the validity of the network model proposed in this study.

Key words: gesture recognition; densely connected 3DCNN; convolutional GRU; spatiotemporal features

1　引言

人们对手势识别技术的研究已有几十年的历程, 经历了不同的发展阶段. 手势识别开始于 1983 年, 来自 AT&T 的 Grimes[1]发明了数据手套, 其通过数据线

① 基金项目: 福建省自然科学基金 (2017J01744)
Foundation item: Natural Science Foundation of Fujian Province (2017J01744)
收稿时间: 2018-05-02; 修改时间: 2018-05-24; 采用时间: 2018-06-05; csa 在线出版时间: 2018-11-28

与计算机相互连接来进行手势定位跟踪和时序信息的检测处理. 采用数据手套的方法数据量小、稳定性和识别准确性高, 但由于需要穿戴昂贵的硬件设备, 操作不方便的同时也对人体进行了限制, 因而难以得到有效的推广, 这也迫使研究者寻求更为自然的方法. 随后的彩色相机的出现, 基于视觉的方式成为主流. 传统的动态手势识别方法主要基于动态时间规整 (DTW)[2]和基于隐马尔可夫模型 (HMM)[3]. 2010 年微软推出的 Kinect 传感器为计算机视觉提供了全新的数据类型, 即深度信息, 它包含着物体到摄像头的距离信息, 深度信息的利用使得视觉处理中较困难的分割过程更为容易, 正是由于可以提供这种有用的深度信息, 使得 RGB-D 相机在手势识别研究被广泛使用.

近年来, 深度学习在图像分类[4]、目标检测[5]、语义分割[6]、场景理解[7]等计算机视觉领域得到广泛使用, 该技术可以对特征进行分层抽象学习, 通过网络训练自动提取特征. 利用深度学习技术进行手势的识别是目前主流的研究方法, 国内外研究人员在各种手势数据集上进行了研究工作. 李宇楠等[8]利用手势 RGB 图像序列及通过 RGB 图像序列计算出的光流序列, 分别使用 3DCNN(3D Convolutional Neural Networks) 网络进行特征提取, 然后对提取的特征进行融合, 利用支持向量机 (SVM) 来进行手势识别; 清华大学的 Chen X 等[9]提出一种运动特征增强的 RNN 网络, 对基于骨架结构的手势序列进行动态手势识别; Molchanov 等[10]等利用 3DCNN 网络对手势时空域进行特征提取, 配合时空特征增强方法, 在 VIVA 数据集上达到 77.5% 的识别率. 目前绝大部分的研究都采用了深度学习技术处理基于视频的手势识别.

本文是对 SKIG RGB-D 多模态的孤立手势视频进行手势识别研究. 对采样出的 32 帧 RGB 图像序列和 Depth 图像序列, 分别利用本文提出的稠密连接的 3DCNN 组件学习短期的时空域特征, 然后将提取的时空域特征输入到卷积 GRU 网络进行长期的时空域特征学习, 最终对单模态训练好的网络进行多模态融合, 提升网络识别准确率. 本文在 SKIG 数据集上取得了 99.07% 的识别准确率.

2 模型架构

基于视频的手势识别涉及到时间和空间因素, 因而不仅要考虑手势的空域特征, 还要考虑时域特征. 对时空域的特征学习是手势乃至其它人体行为识别[11]的重点. LRCN[12]将 CNN 与 LSTM 结合用来提取时空域特征, 先对视频采样出的帧, 通过 CNN 进行空域特征提取, 然后对按序提取出来的空域特征, 利用 LSTM 来学习其时域特征. 双流 CNN 网络利用两条分支分别从 RGB 图像中提取空域特征和堆叠的光流图像中提取时域特征, 对最终的分类进行融合. 这两种具有代表性的方式, 前者采用分阶段学习时空域特征, 而后者是对时空域特征各自独立学习. 考虑到手势背景复杂多变, 对时空域特征同时学习, 是更为有效的方式. 3DCNN 网络就是基于这种理念, 利用三维卷积核对时域和空域同时处理, 这种方式比前两种更为有效, 因而被众多研究者用来对视频进行时空域特征的提取. GRU 对时间序列数据有很好的学习效果, 但是采用全联接的方式, 对空域特征的学习能力较弱. 利用卷积 GRU 网络可以学习长期的时空域特征. 利用本文提出的稠密连接的 3DCNN 学习视频短期的时空域特征, 进而使用卷积 GRU 从短期时空域特征来学习视频长期的时空域特征是合理的组合方式. 本文采用的单模态的网络模型结构见图 1.

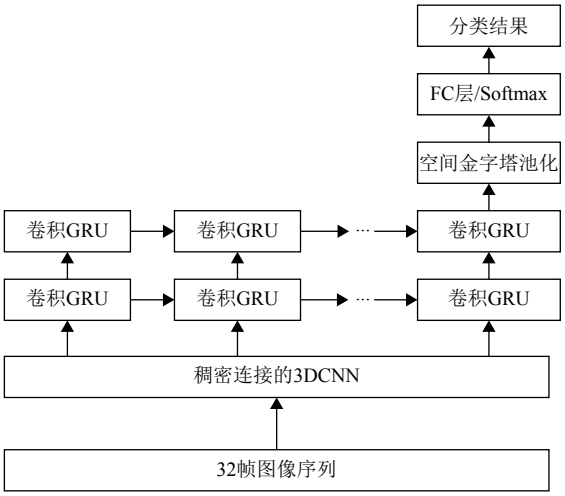


图 1 单模态的网络模型结构

如图 1 所示, 单模态的网络模型结构分为五个部分: (1) 预处理好的 32 帧图像序列, 作为网络的输入部分; (2) 本文提出的稠密连接的 3DCNN 结构, 用于对输入的序列提取短期时空域特征; (3) 双层卷积 GRU 网络, 更进一步对提取的短期时空域特征进行长期时空域特征的学习; (4) 空间金字塔池化层用于降维; (5) 全连接 FC 层的输出使用 Softmax 分类器得到概率向量, 对最终的网络输出进行分类预测. 具体各部分将依次介绍。

2.1 稠密连接的 3DCNN 组件

稠密卷积网络[13](DenseNets) 使用合适的特征尺

寸, 将所有层的特征都进行相互联接, 来获取网络各层间的最大信息, 为了保持前馈性, 每层都对之前的所有层的输出进行拼接后作为本层输入, 得到的输出特征图传递给后续所有层. 依据 DenseNets 网络 Dense block 的思想, 将其应用到 3DCNN, 本文提出稠密连接的 3DCNN 结构用于对手势视频进行短期时空域特征提取. 对提出稠密连接的 3DCNN 结构一些参数的情况加以说明:

(1) 规定网络输入的层的输入图像序列的格式以及特征图的格式按“通道数@长度×高度×宽度”方式标记.

(2)3D 卷积核和 3D 池化核的大小为 $d\times k\times k$, 其中 d 表示时间长度, k 为空间大小. 每个卷积核大小为 3×3×3, 卷积核步长大小均为 1×1×1, Padding 方式选用‘SAME’.

(3)3D 池化核使用是最大值池化.

如图 2 所示的结构中, 输入部分是对视频采样出的 32 帧组成的图像序列. 通过 64 个 3D 卷积核进行卷积操作得到 64@32×112×112 的特征图, 空间尺寸保持不变, 然后利用 1×2×2 池化操作, 保持时间维度不变, 空间尺寸缩小为原来的 1/4. 稠密连接部分每个卷积层的 3D 卷积核个数为 32, 通过跨层拼接的方式, 依次得到的特征图个数为: 32, 64+32=96, 64+32+32=128, 64+32+32+32=160, 然后通过 32 个 3D 卷积核卷积操作, 提取特征后利用 2×2×2 池化进行降维得到 32@16×56×56 的最终输出特征, 作为后续双层卷积 GUR 的输入.

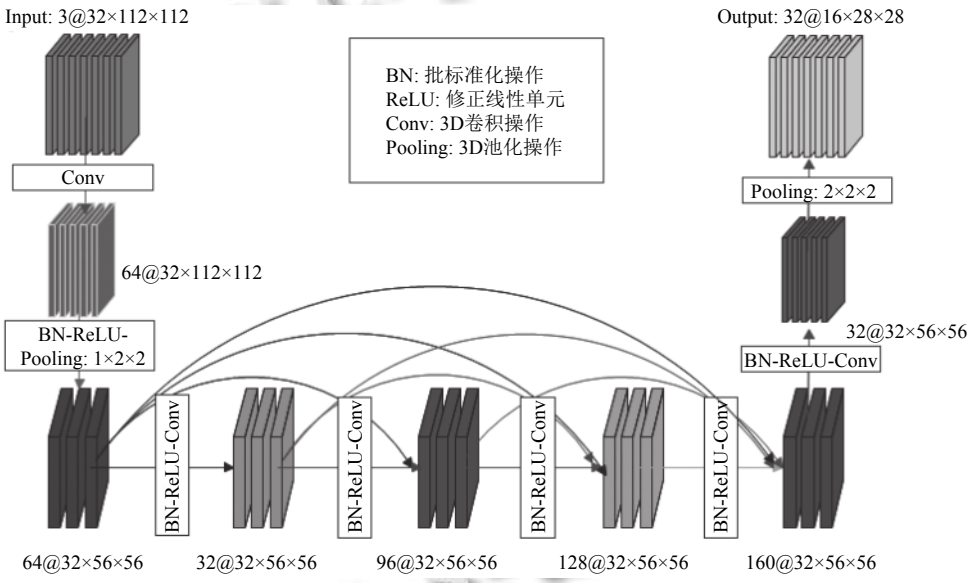


图 2 稠密连接的 3DCNN 结构

2.2 双层卷积 GRU

传统的 GUR 输入到状态, 状态到状态之间的转换是采用全连接的方式, 而全连接方式对空间维度没有进行有效利用, 因而本文使用卷积 GRU, 将全连接方式使用卷积操作代替, 用来对长期的时空域特征同时提取, 具体如公式 (1) 所示:

$$\begin{cases} z_t = \sigma(W_z * x_t + U_z * h_{t-1} + b_z) \\ r_t = \sigma(W_r * x_t + U_r * h_{t-1} + b_r) \\ \bar{h}_t = \tanh(W_h * x_t + U_h * (r_t \circ h_{t-1}) + b_h \\ h_t = (1 - z_t) h_{t-1} + z_t \bar{h}_t \end{cases} \tag{1}$$

其中, $x_1,\cdots,x_t$为不同时刻的输入信息, $h_1,\cdots,h_t$对应不同时刻的隐藏状态, z_t 是更新门, 用来控制当前的状态需要遗忘多少的历史信息和接受多少的新信息, r_t 重置门, 用来控制候选状态中有多少信息是从历史信息中得到, $\bar{h}_t$是候选隐含状态, h_t 是当前时刻的隐含状态, $W*$和$U*$均是 2 维卷积核, σ为 Sigmoid 激活函数, $'\circ'$表示矩阵 Hadamard 积.

本文使用双层的卷积 GRU, 第一层的卷积核数目为 256, 第二层的卷积核数目设为 384, 卷积核的大小均为 3×3, 卷积核步长大小均为 1×1, Padding 方式选用‘SAME’. 将第二层最终学习到的特征作为双层卷积 GUR 的输出, 384@1×28×28, 其中 384 指特征图个数, 28×28 为每个特征图的空间大小, 时间长度为 1.

2.3 空间金字塔池化层

双层卷积 GRU 输出为 384@1×28×28, 总的维度太高, 要先进行降维处理, 本文使用了 4 种层次的 SPP, 分别是 1×1、2×2、4×4、7×7 结构, 如图 3 所示, 最终生成 1+4+16+49=70 个 384 维的特征, Flatten 变平化为 1 维向量后的结果为 1×70×384=26880, 再与全连接层相连. 采用多层 SPP 降维的同时对同一特征图多种尺度的提取特征, 对网络识别精度有所提高.

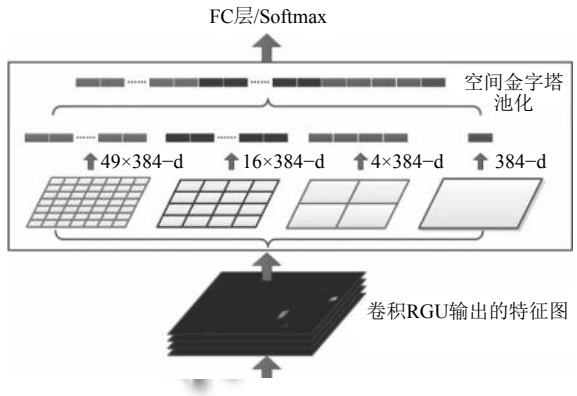


图 3 空间金字塔池化层

2.4 模型融合

多模态融合是常用的提升模型准确度的方法, 本文融合模型是对训练好的两种模态网络的 Softmax 层输出的概率向量进行相加除以 2, 选取最终得到的融合概率向量中数值最大的概率所对应的类别作为分类的结果, 融合模型如图 4 所示.

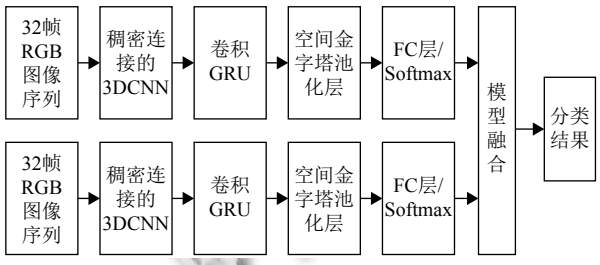


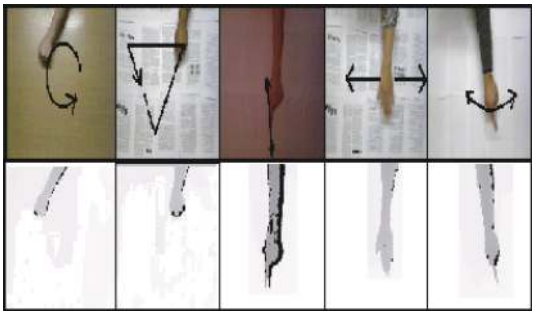
图 4 多种模态的融合模型结构

3 实验验证及结果分析

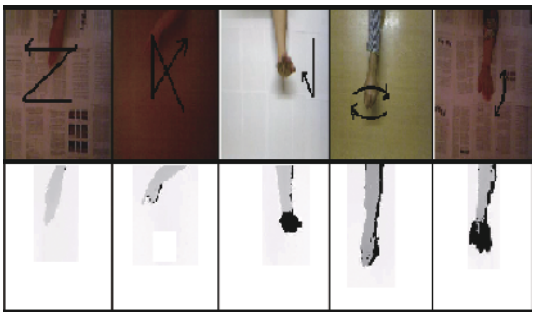
3.1 数据集

本文基于 Sheffield Kinect Gesture (SKIG) Dataset[14] RGB-D 孤立手势视频数据集, 对提出的手势识别网络模型进行实验, 数据集类别共 10 类, 如图 5 所示, 图中展示了 RGB 图像及所对应的 Depth 图像.

SKIG 数据集包含手势的 RGB 视频及 Depth 视频两种模态, 该手势数据集是利用微软 Kinect 设备的 RGB 摄像头和深度摄像头, 同步采集人体手势而得到, 数据集没有划分训练集与测试集. 具体细节如下:



(a) SKIG 前 5 种手势类别



(b) SKIG 后 5 种手势类别

图 5 SKIG 前后 5 种手势类别

(1) 一共采集了 6 人 (subject) 的手势, 每个手势的 RGB 视频有相应的 Depth 视频. (2) 包含 10 个手势类别: Circle(画圆)、Triangle(画三角形)、Up-down(上下移动)、Right-left(右左移动)、Wave(挥手)、‘Z’(画 Z 字形)、Cross(画十字形)、Come here(招唤动作)、Turn around(翻转) 以及 Pat(轻拍). (3) 每种手势分别使用 3 种手形执行: 握拳、伸食指和张开手掌. (4) 采用 3 种背景: 木板、白纸和报纸. (5)2 种光照: 较亮和较暗. (5) 总视频数 2160, RGB 视频和 Depth 视频各占一半 (6×10×3×3×2=1080 个).

3.2 实验环境

(1) 硬件环境: NVIDIA Tesla P40 24 GB 显卡 8 核 32 GB CPU

(2) 软件环境: CentOS7 操作系统 Python 3.5.2 版 TensorFlow 1.2.1 版 TensorLayer 1.6.5 版 CUDA8.0 cuDNN5.0

3.3 模型参数

因为实验用到的网络模型是第一次提出, 整个网络从头开始训练, RGB 模态和 Depth 模态数据集各自

独立训练, 两种模态的网络训练参数设为一致. 批次大小为 18; 学习率初值设为 0.001; 权重衰减系数设为 0.0004; 每 6000 次迭代, 学习率下降为原来的 1/10; 网络训练时每迭代 500 个批次, 就对测试集进行一次测试; 训练的周期数, 设为 300 个周期, 对应 12000 左右的迭代次数.

3.4 实验及结果分析

数据集没有划分训练集与测试集, 采用文献[15]中的 3 折交叉验证, 将 6 个 subjects, 划分成三个子集, 其中子集 1 为: subject1+subject2; 子集 2 为: subject3+subject4; 子集 3 为: subject5+suject6.

分组 1: 训练集为子集 1 和子集 2, 测试集为子集 3, 结果如图 6 所示, 经测试选取的两个训练好的单模态网络模型参数为: RGB 数据集 11 000 次迭代时测试准确度为 98.33% 的模型参数和 Depth 数据集 10 000 次迭代时测试准确度为 99.17% 的模型参数.

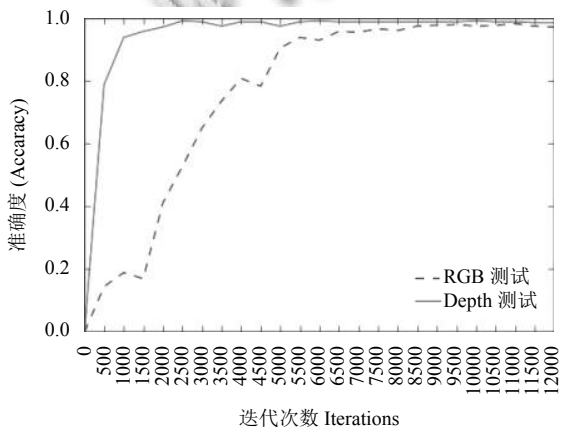


图 6 分组 1 的测试结果

分组 2: 训练集为子集 1 和子集 3, 测试集为子集 2, 结果如图 7 所示, 经测试选取两个训练好的单模态网络模型参数为: RGB 数据集 10 000 次迭代时测试准确度为 96.94 % 的模型参数和 Depth 数据集 10 500 迭代时准确度为 97.78 % 的模型参数.

分组 3: 训练集为子集 2 和子集 3, 测试集为子集 1, 结果如图 8 所示, 经测试选取的最优的两个训练好的单模态网络模型参数为: RGB 数据集 11500 次迭代时准确度为 93.06% 的模型参数和 Depth 数据集 9000 迭代时准确度为 99.17 % 的模型参数.

对每个分组单模态网络各自训练好的模型, 按本文所用的方法进行模型融合, 得到各分组多模态融合后的准确率, 如表 1 所示.

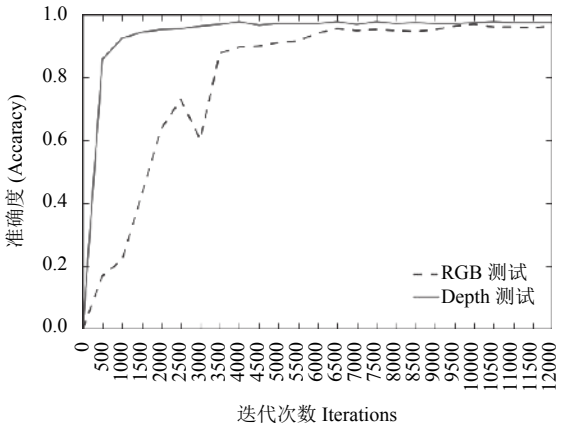


图 7 分组 2 的测试结果

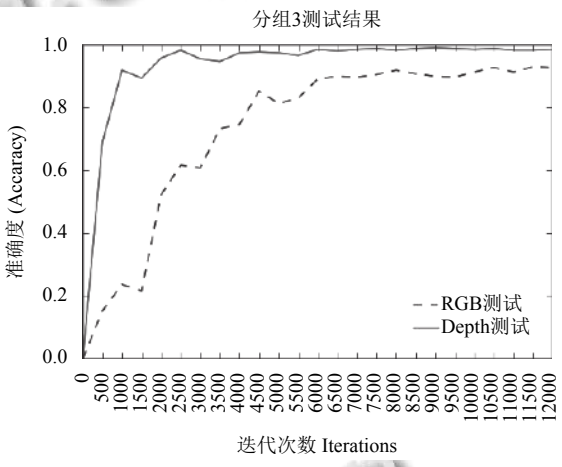


图 8 分组 3 的测试结果

表 1 各模态的实验准确率 (%)

	RGB 模态	Depth 模态	模态融合
分组 1	98.33	99.17	99.44
分组 2	96.94	97.78	98.06
分组 3	93.06	99.17	99.72
平均值	**96.11**	**98.71**	**99.07**

将本文方法结果与近几年在 SKIG 数据集上相关实验的结果进行对比, 如表 2 所示, 本文提出的方法具有更高的准确率, 达到 99.07%. 其中 RGGP+RGB-D 方法使用受限图形遗传编程 (RGGP) 方法, 从视频中自动提取具有鉴别性的时空特征, 对 RGB 和 Depth 信息的融合来进分类, 识别率为 88.7%, 与本文准确率相差 10.37%. MRNN 方法利用 2DCNN 对视频的空间特征进行学习, 学习到的特征输入到 MRNN 网络进行手势分类, 与本文准确率差了 1.27%. 3DCNN+CLSTM 利用 3DCNN 结合 CLSTM 的方法来进行时空域的学习, 达到了 98.89% 的准确率, 它使用的是传统的 3DCNN,

与本文提出的稠密连接的 3DCNN 在特征的处理上并不相同, 本文的模型参数少于其一半, 约 930 万, 大幅降低模型参数的同时保持相对应的性能, 本文模型提升了约 0.2%.

表 2 不同方法在 SKIG 上的比较

方法	准确率 (%)
RGGP+RGB-D[14]	88.7
MRNN[15]	97.8
3DCNN+CLSTM [16]	98.89
本文模型	99.07

4 结语

本文提出的稠密连接的 3DCNN 结构, 实现对多层特征图进行重复利用, 使得参数利用效率更高, 更加容易进行网络的训练. 通过对不同层的特征进行稠密的组合, 可以对后续层的输入增强多样性, 在提升网络的性能的同时, 降低网络模型的参数量. 利用卷积 GRU 相比传统的 GRU 而言增加了对空间信息的处理能力, 因而能更好的对长期时空域特征进行提取. 本文模型参数及卷积核个数的设置并不是最优, 双向卷积 GRU 可能会进一步提升模型准确率. 后续计划将注意力机制引入, 期望有更好的性能提升.

参考文献

1 Grimes GJ. Digital data entry glove interface device. US, US 4414537. 1983-11-08.

2 Hartmann B, Link N. Gesture recognition with inertial sensors and optimized DTW prototypes. Proceedings of 2010 IEEE International Conference on Systems, Man and Cybernetics. Istanbul, Turkey. 2010. 2102–2109.

3 Liu NJ, Lovell BC, Kootsookos PJ, *et al*. Model structure selection & training algorithms for an HMM gesture recognition system. Proceedings of 9th International Workshop on Frontiers in Handwriting Recognition. Kokubunji, Tokyo, Japan. 2004. 100-105.

4 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

5 He KM, Gkioxari G, Dollár P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2980–2988.

6 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.

7 Yu TS, Wang RS. Enhancing scene parsing by transferring structures via efficient low-rank graph matching. Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Burlingame, CA, USA. 2016.

8 Li YN, Miao QG, Tian K, *et al*. Large-scale gesture recognition with a fusion of RGB-D data based on the C3D model. Proceedings of the 23rd International Conference on Pattern Recognition. Cancun, Mexico. 2017. 25–30.

9 Chen XH, Guo HK, Wang GJ, *et al*. Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. Proceedings of 2017 IEEE International Conference on Image Processing. Beijing, China. 2017.

10 Molchanov P, Gupta S, Kim K, *et al*. Hand gesture recognition with 3D convolutional neural networks. Proceedings of 2015 IEEE Computer Vision and Pattern Recognition Workshops. Boston, MA, USA. 2015. 1–7.

11 Ding M, Fan GL, Zhang X, *et al*. Structure-guided manifold learning for video-based motion estimation. Proceedings of the 19th IEEE International Conference on Image Processing. Orlando, FL, USA. 2012. 1977–1980.

12 Donahue J, Hendricks LA, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of 2015 IEEE Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 677–691.

13 Huang G, Liu Z, van der Maaten L, *et al*. Densely connected convolutional networks. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017.

14 Liu L, Shao L. Learning discriminative representations from RGB-D video data. Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China. 2013. 1493–1500.

15 Nishida N, Nakayama H. Multimodal gesture recognition using multi-stream recurrent neural network. Proceedings of the 7th Pacific-Rim Symposium on Image and Video Technology. Auckland, New Zealand. 2015.

16 Zhu GM, Zhang L, Shen PY, *et al*. Multimodal gesture recognition using 3-D convolution and convolutional LSTM. IEEE Access, 2017, 5: 4517–4524. [doi: 10.1109/ACCESS.2017.2684186]