文章编号:1671-4598(2022)02-0276-08 DOI:10.16526/j.cnki.11—4762/tp.2022.02.040 中图分类号:TP391.41 文献标识码:A

基于注意力机制的多模态人体行为识别算法

宋真东[1,2]，杨国超[1]，马玉鹏[1,3]，冯晓毅[1]
(1. 西北工业大学 电子信息学院，西安 710129；
2. 陕西华明普泰医疗设备有限公司，西安 710119)
3. 河北师范大学 计算机与网络空间安全学院，石家庄 050024；

摘要：提出了基于注意力机制的多模态人体行为识别算法；针对多模态特征的有效融合问题，设计基于注意力机制的双流特征融合卷积网络（TAM3DNet，two—stream attention mechanism 3D network)；主干网络采用结合注意力机制的注意力3D网络（AM3DNet，attention mechanism 3D network)，将特征图与注意力图进行加权后得到加权行为特征，从而使网络聚焦于肢体运动区域的特征，减弱背景和肢体静止区域的影响；将RGB—D数据的颜色和深度两种模态数据分别作为双流网络的输入，从两条分支网络得到彩色和深度行为特征，然后将融合特征进行分类得到人体行为识别结果。

关键词：RGB—D图像；多模态特征；人体行为；双流网络；注意力机制；特征融合

Multi—modal Human Behavior Recognition Algorithm Based on Attention Mechanism

SONG Zhendong[1,2]，YANG Guochao[1]，MA Yupeng[1,3]，FENG Xiaoyi[1]
(1. School of Electronics and Information，Northwestern Polytechnical University，Xi'an 710129，China；
2. Shanxi Huaming Putai Medical Equipment，Xi'an 710119，China；
3. College of Computer and Cyber Security，Hebei Normal University，Shijiazhuang 050024，China)

Abstract：A multi—modal human behavior recognition algorithm based on the attention mechanism is proposed. Aiming at effective fusion problem of the multimodal features，a two—stream feature fusion convolutional two—stream attention mechanism 3D network (TAM3Dnet) based on the attention mechanism is designed. The backbone network adopts the attention mechanism 3D network (AM3DNet)，which combines with the attention mechanism，and weights the feature and attention map to obtain the weighted behavior characteristics，so that the network focuses on the characteristics of the limb movement area and reduces the influence of the background and limb rest area. The color and depth modal data for the RGB—D are respectively used as the input of the dual—stream network，and the color and depth behavior features are obtained from two branch networks，and then the fusion features are classified to obtain the human behavior recognition results.

Keywords：RGB—D image；multi—modal features；human behavior；two—stream network；attention mechanism；feature fusion

0 引言

传统的行为识别方法使用普通的RGB数据来进行，但难以有效解决光照变化、背景复杂、遮挡等因素影响。近年来出现了许多方便操作、价格便宜的多

收稿日期:2022-01-07； **修回日期**:2022-01-11。
基金项目:陕西省重点研发项目(2021ZDLGY15—01，2021ZDLGY09—04，2021GY—004和2020GY—050)；深圳市国际合作研究项目(GJHZ20200731095204013)；国家自然基金(61772419)。
作者简介:宋真东(1968-)，男，江苏泰县人，硕士研究生，主要从事机器视觉、行为识别方向的研究。
通讯作者:马玉鹏(1987-)，男，河北宁晋人，博士，主要从事计算机视觉、行为识别方向的研究。
引用格式:宋真东，杨国超，马玉鹏，等.基于注意力机制的多模态人体行为识别算法[J].计算机测量与控制，2022，30(2)：276-283.

模态摄像机，通过彩色深度传感器（RGB－D，Red、Green、Blue 和 Depth）[1]可以同时采集 RGB 图像和 Depth 图像，能够提供彩色图像不具备的三维运动和结构信息，为提高行为识别系统的鲁棒性和准确性提供有效支撑。因此，近年来基于 RGB－D 多模态数据的人体行为识别引起关注。

深度学习在语言处理、计算机视觉和视频理解等领域已有广泛深入的应用。K. Simonyan 等人[2]提出的 Two－Stream 双流网络是深度学习的一个主流方向，该算法使用两个并行的网络分支分别学习视频的空间特征和时间特征，以单帧的 RGB 图像输入网络提取空间场景和目标信息，将密集光流序列输入网络来学习时间特征，最后将两个分支的判断进行融合得到分类结果。C. Feichtenhofer 等人[3]在 Two－Stream 网络的基础上利用 CNN 网络进行时空特征融合，并将基础网络替换成 VGG－16，提高了识别效果。Z. Liu 等人[4]提出了 3D 卷积神经网络（3DCNN，3D－based deep convolutional neural network），3 维卷积核相比 2 维卷积核多了一个时间维度，因此该网络可以自动地学习时空特征，视频描述子具有高效通用的特点。W. Du 等人[5]将长短期记忆网络（LSTM，long short－term memory）[6]与 CNN 结合提出了循环姿势注意力网络（RPAN，recurrent pose－attention network）算法，该算法包括特征生成、姿态注意机制和 LSTM 时序网络三部分，LSTM 解决了一般的循环神经网络（RNN，recurrent neural networks）[7]依赖前后长期信息的问题，适合提取时间维度特征。

现有的行为识别方法主要是对视频帧整体提取特征，没有区分行为感兴趣区域和静止区域，且很多方法仅利用 RGB 单模态信息，因此，行为识别准确性难以满足实际需求。针对面向行为识别的区域检测问题，本文借鉴生物视觉系统的注意力机制，结合 3D 卷积网络构建了基于注意力机制的 3D 卷积网络（AM3DNet，attention mechanism 3D network），能有效提取与行为识别相关的肢体运动部位特征。针对 RGB 图像和 Depth 图像多模态输入及特征融合问题，提出了基于注意力机制的 RGB－D 双流特征融合 3D 网络（TAM3DNet，two－stream attention mechanism RGB－D feature fusion 3D network），RGB 图像和 Depth 图像作为双流网络的输入，主干网络采用 AM3DNet 分别提取 RGB 图像特征和 Depth 图像特征，再将融合后的特征输入网络分类层，得到最终的行为识别结果。

1　3D 卷积和注意力机制

1.1　3D 卷积

2D 卷积提取单张静止图像的空间特征，适用于图像的分类、检测等任务。2D 卷积在行为识别任务中对每一帧图像分别提取空间特征，一个卷积核只能得到一个特征图，这种卷积方式没有考虑时间维度帧间的物体运动信息，因此，2D 卷积不适用于视频和多帧图像等具有时间维度信息的任务。

为了提取视频数据的时间维度特征，提出了 3D 卷积。3D 卷积在卷积核中加入了时间维度，能同时提取视频帧的空间和时间特征信息[8]。3D 卷积与 2D 卷积的不同之处在于，输入的数据和卷积核都增加了一个维度，多个连续的视频帧组成一个立方体作为输入，然后在立方体中运用 3D 卷积核，卷积层中的每一个特征图都是从上一层中多个连续帧提取得到。因此，3D 卷积能捕捉到运动信息，适用于行为识别任务。2D 卷积和 3D 卷积操作如图 1 所示。

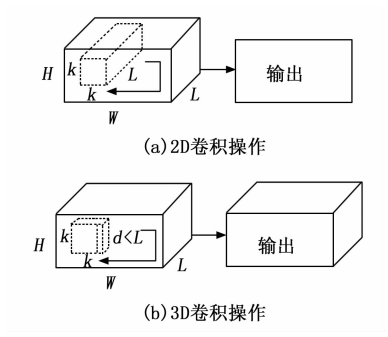

图 1

1.2　注意力机制

生物视觉系统通常不会关注场景中的所有区域，而是关注场景中的关键位置来获取有用信息，这就是生物视觉的注意力机制[9-10]。基于注意力机制的模型在深度学习的各个领域中广泛应用，可有效提高深度学习任务的性能。基于注意力机制的模型，首先快速扫描全局图像得到重点关注的目标区域，然后对这一区域集中注意力资源获得更多关注目标的细节信息，抑制周围的无关信息，极大提高了视觉信息处理的效率和准确度。

近年来，深度学习与注意力机制结合的研究主要

集中在使用掩码（mask）来实现。掩码的原理是通过一层新的权重，标识出图像中关键的特征，通过训练使神经网络学习每张图像中需要重点关注的区域，从而实现注意力机制。这种方式演化为两种类型的注意力，一种是软注意力（soft attention），另一种是强注意力（hard attention），以下分别介绍：

1）软注意力：软注意力是确定性的注意力，更加关注区域[11]或通道，学习完成后可以直接通过网络生成权重，保留所有特征分量进行加权。最重要的一点是软注意力是可微分的，首先可微分的注意力可以通过神经网络计算出梯度，然后梯度下降法通过目标函数及相应的优化函数来学习注意力权重。

2）强注意力：与软注意力不同，强注意力更加关注像素点[12]，图像中每个点都可能得到注意力，而且强注意力更加强调动态变化，是一个随机预测的过程，选取部分特征进行加权。最关键的是强注意力是不可导的注意力，往往通过强化学习（reinforcement learning）来完成训练，强化学习通过收益函数（reward）来激励，使模型关注局部的细节信息。

2 模型与方法

行为识别的关键问题在于准确提取感兴趣行为特征和多模态特征的有效融合，目前行为识别方法对图像整体提取特征，没有区分肢体运动区域和其它区域[13]，本文将注意力机制和 3D 卷积网络相结合，使肢体运动部位的特征作为重点。针对 RGBD 多模态特征有效融合问题，通过实验对比选择特征层拼接融合方式，借鉴双流网络结构，用深度图代替光流图，提出基于注意力机制的双流特征融合卷积网络 TA3D。

2.1 基于注意力机制的 3D 卷积网络

视觉注意力机制本质是在图像的不同区域加上不同权重，使用注意力机制有利于提高行为识别判断的准确性。常规的 3D 卷积网络对视频帧所有空间区域的作用是一致的，不能区分运动区域和非运动区域。本文提出的结合注意力机制的 3D 卷积网络对模型的学习能力进行分配，使图像中与行为识别相关的区域权重增大，降低无关区域的权重。视觉注意力模块如图 2 所示。

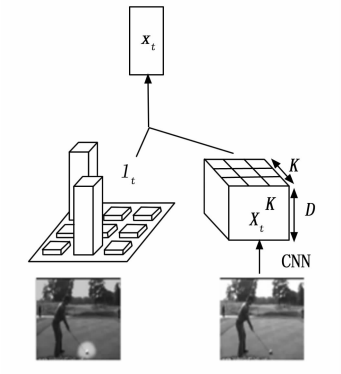

图 2 视觉注意力模块

其中：X_t 表示第 t 帧视频帧通过 CNN 卷积网络后得到的特征图，尺寸为 $K\times K\times C$，其中 K 代表特征图的空间大小，C 代表特征图的通道维度。l_t 表示对应于第 t 帧的注意力图，其为 $K\times K$ 的向量。将注意力图和卷积图通过加权相结合后得到输出 x_t，如式（1）所示，然后将 x_t 输入到后续网络中，得到的输出经过非线性变换后作为网络当前时刻的输出。

$$x_t = E_{p(L_t|h_{t-1})}[X_t] = \sum_{i=1}^{K^2} l_{t,i} X_{t,i} \qquad (1)$$

式中，X_t 是 t 时刻的特征立方体，$X_{t,i}$ 是 t 时刻特征立方体的第 i 个切片。$l_{t,i}$ 是 t 时刻注意力图的第 i 个权值向量，得到的 x_t 是大小为 C 的特征向量，其中 C 是特征图的通道维度。卷积神经网络输出的特征图尺寸为 $K\times K\times C$，如果沿着特征图的空间维度展开，可以当成是 $K\times K$ 个 d 维的向量，相当于将特征图分块表示，每个向量对应输入视频帧不同区域的特征值。如图 1 所示，为了与特征图相结合，注意力图的尺寸应该与特征图空间尺寸一致，且注意力图不同部位的向量大小表示对应特征图区域的权重大小。经过加权运算后，加强运动区域的卷积特征，减弱背景和静止区域的卷积特征。

由于注意力机制在计算机视觉领域特别是视频分类识别方面具有较大优势，本文将注意力机制（AM，attention mechanism）与原始 3D 卷积网络相结合：在 3D 卷积层后加入注意力层，使用自注意力机制计算注意力图，其余网络结构不变，如图 3 所示。本文将改进后的网络命名为注意力机制 3D 卷积网络（AM3DNet，attention mechanism 3D network），该网络首先通过 3D 卷积层提取视频帧序列的特征图，然后将视频帧序列特征图输入注意力模块获得当前输入的注意力图，之后将序列特征图和与之

对应的注意力图加权融合后得到加权特征，从而加强对当前行为识别任务重要的肢体运动区域特征并且抑制不重要的区域特征，再将加权后的特征输入后续3D卷积层和全连接层，最后通过Softmax层得到行为类别预测结果。该网络通过学习特征空间不同区域的权重分布，使网络专注于对行为识别有意义的肢体运动部位，可提高行为识别网络的性能。

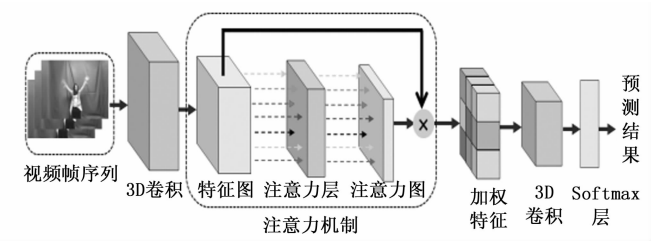

图 3　AM3DNet 结构示意图

注意力图是由嵌入在网络中的注意力层得到，目的是动态地估计不同视频帧之间的显著性和相关性[14]，假设视频帧经过前端3D卷积层后得到的特征图F尺寸为$K\times K\times C$，C为通道数。注意力层是1×1×1的3D卷积核，在特征图（i，j）处的单位立方体F_{ij}内进行卷积得到值A_{ij}，该值代表原始视频帧中对应区域的权重，所有区域的权重值组合为一个尺寸与特征图相同的矩阵$\boldsymbol{A}$，区域注意力权重A_{ij}的计算如式（2）所示：

$$A_{ij}=Sigmoid(W_{ij}F_{ij}+b_{ij}) \tag{2}$$

式中，W_{ij}是变换矩阵，F_{ij}是（i，j）处的特征向量，b_{ij}是偏置项，使用Sigmoid函数作为激活函数将注意力权重约束在[0,1]区间内，注意力权重矩阵$\boldsymbol{A}$与特征图F逐项相乘后得到加权特征图，然后输入后续网络进行特征提取和分类。该网络使用的损失函数如式（3）所示：

$$L=-\sum_{t=1}^{T}\sum_{i=1}^{C}y_{t,i}\log\hat{y}_{t,i}+\lambda\sum_{i=1}^{K^2}(1-\sum_{t=1}^{T}l_{t,i})^2 \tag{3}$$

式中，第一项为交叉熵损失函数，是分类问题中常用的损失函数，其中y_t是数据标签向量，是t时刻的类别概率向量，T代表总的时间步数，C代表输出的类别数。第二项为随机惩罚项，λ是注意力惩罚系数，括号内是视频帧中第i个区域对应注意力图的权重值，其在所有区域内的和为1。

2.2　RGB－D双流网络的融合方式

多模态数据的网络融合方式主要分为特征层融合和决策层融合[15-16]。其中，特征层融合是指多个网络分支学习的特征融合在一起，然后将融合后的特征输入分类器得到分类结果。决策层融合是指在预测级别进行融合，多个独立网络训练后得到不同模型，测试时每个模型都会得到预测分数，将预测分数进行融合后得到最终的预测结果。

本文通过实验选择特征层融合，即首先将RGB图像和Depth图像分别输入网络中，获得RGB图像的特征与Depth图像的特征；然后两种特征在通道维度上进行融合，得到融合后特征；最后将融合后特征输入分类器中得到预测结果。特征层融合机制如图4所示。

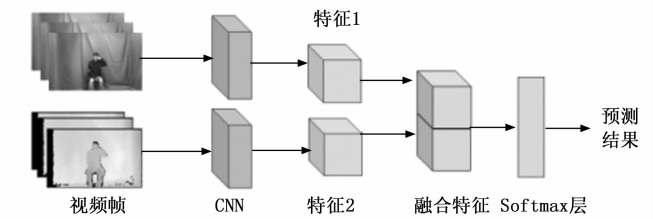

图 4　特征层融合机制

2.3　基于注意力机制的RGB－D双流特征融合3D网络

RGB－D图像两种模态作为输入，借鉴Two－Stream网络的结构[17]，本文在AM3DNet的基础上提出了基于注意力机制的RGB－D双流特征融合3D网络（TAM3DNet，two－stream attention mechansim RGB－D feature fusion 3D network），其为结合注意力的双流特征融合网络，TA3D的结构如图5所示。首先将RGB数据和Depth数据预处理后作为双流网络两条流的输入，主干网络使用结合注意力机制

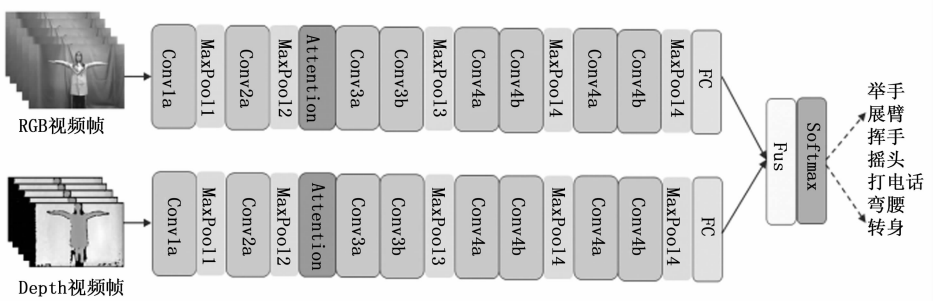

图 5　TAM3DNet 结构示意图

表 1 TAM3DNet 模型参数

层	kernel num	kernel size	stride	RGB 网络尺寸	Depth 网络尺寸
Input	—	—	—	[16, 112, 112, 3]	[16, 112, 112, 1]
Conv1	64	[3,3,3]	[1,1,1]	[16, 112, 112, 64]	[16, 112, 112, 64]
Pool1	—	[1,2,2]	[1,2,2]	[16, 56, 56, 64]	[16, 56, 56, 64]
Conv2	128	[3,3,3]	[1,1,1]	[16, 56, 56, 128]	[16, 56, 56, 128]
Pool2	—	[2,2,2]	[2,2,2]	[8, 28, 28, 128]	[8, 28, 28, 128]
Attention	1	[1,1,1]	[1,1,1]	[8, 28, 28, 1]	[8, 28, 28, 1]
Conv3	256	[3,3,3]	[1,1,1]	[8, 28, 28, 256]	[8, 28, 28, 256]
Pool3	—	[2,2,2]	[2,2,2]	[4, 14, 14, 256]	[4, 14, 14, 256]
Conv4	512	[3,3,3]	[1,1,1]	[4, 14, 14, 512]	[4, 14, 14, 512]
Pool4	—	[2,2,2]	[2,2,2]	[2, 7, 7, 512]	[2, 7, 7, 512]
Conv5	512	[3,3,3]	[1,1,1]	[2, 7, 7, 512]	[2, 7, 7, 512]
Pool5	—	[2,2,2]	[2,2,2]	[1, 4, 4, 512]	[1, 4, 4, 512]
Fc	4096	—	—	[4096]	[4096]

的 AM3D 卷积网络，将注意力层嵌入卷积层后，分别提取两类数据的注意力加权特征。TAM3D 网络选择特征拼接方式将 RGB 和 Depth 图像的注意力加权特征进行融合，最后将融合特征输入分类层得到分类结果。

在深度学习中需要评估标签值 label 和预测值 predicts 之间的差距，常使用交叉熵作为损失函数来评价模型。假设第 i 类的标签值为 y_i，经过 Softmax 层输出的预测概率为 $\hat{y}_i$，则交叉熵损失函数如公式（4）所示：

$$Loss = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{4}$$

如果网络是批量输入的，假设 batch 的样本数为 m，则对应于一个 batch 批量的整体损失 $loss$ 计算如式（5）所示：

$$Loss = \frac{-1}{m} \sum_{j=1}^{m} \sum_{i=1}^{n} y_{ji} \log(\hat{y}_{ji}) \tag{5}$$

本文提出的 TAM3DNet 分别在双流网络的两个分支中计算各自的交叉熵，然后将两类交叉熵损失之和作为 TAM3DNet 整体的损失函数，针对该损失函数采用 Adagrad 优化器进行优化，寻找损失之和尽可能小的最优参数值。

基于注意力机制的双流特征融合卷积网络 TAM3DNet 参数如表 1 所示。

3 实验结果与分析

3.1 数据集

3.1.1 MSR DailyAction3D 数据集

MSR DailyAction 3D（MSRDA）日常行为数据集是由微软的 Wang 等人[18]在雷德蒙研究院建立，该数据集由 10 个不同的人执行 16 类日常行为动作得到。16 类行为分别为：喝水、吃东西、读书、打电话、写字、欢呼、静坐、使用笔记本电脑、使用吸尘器、走路、弹吉他、扔纸、打游戏、躺在沙发上、站起来、坐下，该数据集记录了执行每个动作的 RGB 视频，以及动作对应的 Depth 图像和 20 个骨架节点的空间位置信息。该数据集每种模态包括 10×2×16＝320 个样本，数据集的 3 种模态总共有 960 个样本。

3.1.2 NPUAction 自建数据集

NPUAction 数据集由 16 个人执行 7 类运动相关动作得到，包括：举手、展臂、挥手、摇头、打电话、弯腰、转身。3D 传感摄像头拍摄得到 RGB 视频片段，同时将 Depth 图像保存为 oni 格式。由于拍摄得到的是连续执行 7 类动作的整段视频，不符合深度学习数据按类别存放的要求，所以人工将整段视频按照行为类别剪辑为 7 段短视频，每段时长在 10 秒钟左右，并按照类别和人物的顺序依次命名，总共得到 224 段 RGB 视频样本。

3.2 实验环境

由于视频处理问题需要大量的计算资源，本文选择在性能强大的 Linux 系统上运行，版本为 Ubuntu 18.04 LTS，运行环境为 Python3.6，使用 RTX 2070 显卡进行运算，CUDA9.0 并行计算架构能加快运算速度，开发工具为 Visual Studio Code，深度学习框架为 GPU 版本的 Tensorflow 1.8.0。

3.3　与主流方法对比实验及结果

为了比较本文提出的基于注意力机制的 RGB－D 双流特征融合 3D 卷积网络 TAM3DNet 与目前主流行为识别算法的性能，在公开的 RGBD 数据集 MSR DailyAction 3D 日常行为数据集和自制 NPUAction 数据集上进行实验。

3.3.1　MSR DailyAction 3D 数据集

在 MSR DailyAction3D 数据集上训练 TA3D 网络模型，对测试集进行多次测试并取准确率平均值，实验结果与改进密集轨迹算法 iDT[19] 和时间段网络 TSN[20] 的准确率对比如表 2 所示。

表 2　MSR DailyAction 3D 数据集上实验结果

行为识别方法	准确率/%
iDT[19]	85.94
TSN[20]	89.06
TA3D	92.19

由实验结果表 2 可以看出，本文提出的 TAM3DNet 在公开的 MSR DailyAction3D 日常行为数据集上取得了 92.19%的识别准确率，与传统算法 iDT 相比识别准确率提高 6.25%，与深度学习算法 TSN 相比提高 3.13%，该结果说明本文提出的基于注意力机制的 RGB－D 双流特征融合 3D 卷积网络在 RGBD 数据行为识别问题上达到了目前先进水平。在 MSRDailyAction 3D 数据集上训练过程的特征图如图 6 所示。

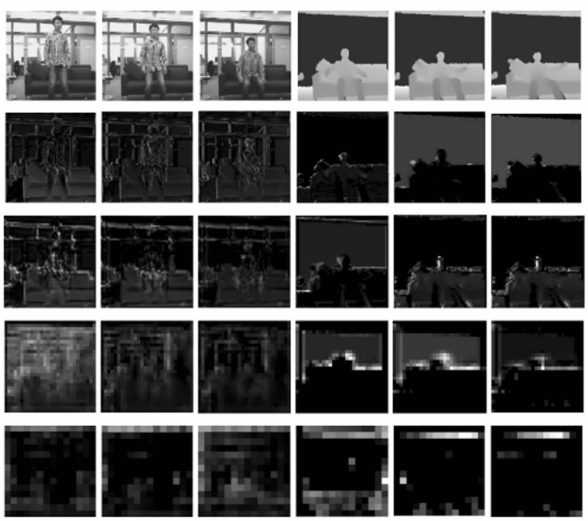

图 6　可视化训练特征图

3.3.2　NPUAction 数据集

为了证明本文提出的基于注意力机制的双流特征融合卷积网络 TAM3DNet 在智慧客厅场景中的识别效果，使用 NPUAction 数据集进行实验，得到整体准确率和每种类别准确率如图 7 所示。

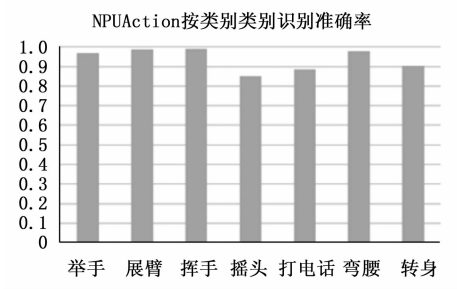

图 7　NPUAction 按类别的识别准确率

由实验结果总结得到，本文提出的 TAM3DNet 在自建场景数据集 NPUAction 上的整体识别准确率达到了 94.05%，由于公开数据集是在普通室内场景中采集的，视频背景光照情况复杂，存在人物遮挡影响，自建 RGB－D 数据集是在实验室模拟环境下采集的，光照和角度可控，更符合本文研究的智慧客厅场景。由图 5 可以看出不同行为的识别准确率差别较大，动作幅度较小的摇头、打电话等动作准确率较低，幅度较大的举手、弯腰等动作识别准确率较高，这个结果也符合视觉系统容易识别大幅度动作的机理，同时也验证了肢体运动部位对行为识别的重要性。实验结果表明，本文提出的 TAM3DNet 在智慧客厅场景中是一个高效的 RGBD 多模态数据端到端行为识别网络。

3.4　注意力机制实验及分析

计算机视觉中的注意力机制是赋予神经网络" 注意力" 能力，使其能集中与图像重点区域而忽略无关信息。为验证注意力机制对人体行为识别所带来的性能提升，在 MSR DailyAction 3D 数据集和 NPUAction 数据上，分别针对三通道 RGB 图像、四通道 RGBD 图像以及 RGB－D 双流特征融合网络进行消融实验。

从实验结果可以看出，基于注意力机制的 RGB－D 双流特征融合网络 TAM3DNet 获得了最好识别结果。在三通道、四通道和 RGB－D 双流输入中，通过增加注意力机制（AM）均能提升识别结果的准确率，从而验证了注意力机制的有效性。四通道 RGBD 输入是由 Depth 图像与 RGB 图像拼接组成，RGB 图像与 Depth 图像是由两个摄像头独立采集得

到，成像原理、帧率以及保存格式据不相同，即便是在融合时进行归一化处理，仍不能保证两种图像准确对齐，导致拼接融合后的识别结果低于三通道数据。同时也说明了 RGB 图像与 Depth 图像的双流特征融合方式的可靠性。

表 3 注意力机制的消融实验

网络模型	NPUA 平均准确率/%	MSRDA 平均准确率/%
三通道 3DNet	81.25	75
三通道 AM3DNet	88.75	83.75
四通道 3DNet	52.5	48.43
四通道 AM3DNet	68.75	57.81
T3DNet	87.45	85.64
TAM3DNet	94.05	92.19

3.5 RGB－D 双流网络融合方式实验及分析

多模态的融合方式有特征层融合和决策层融合两种。为了对比决策层融合与特征层融合方式的优劣，使用 NPUAction 数据集进行两种融合方式对比实验，同时输入 RGB 图像与 Depth 图像，首先分别对两类数据预处理，获取所有视频帧文件的索引，并以 4∶1 的比例划分为训练集和测试集，索引中每行文件的类别要保持一致，才能保证每次输入两条流的数据是同一行为类别的数据，对网络进行有效训练。clip length 取 16，即每次从文件中抽取 16 个视频帧作为一个样本输入模型。

由于双流模型的数据量相对于单流模型大大增加了，限于计算机的硬件条件，本实验将 batch _ size 设置为 2，即每次为训练和测试从硬盘上读取 2 个视频文件，每个视频取 16 帧图像，组成 2 个 clips 作为每条流网络的输入。NPUAction 数据集共有 7 类行为，将 num _ class 设置为 7，每帧统一裁剪为 112×112 的大小。RGB 数据的通道数设置为 3，Depth 数据的通道数设置为 1，对应的网络通道数也作出相应改变。初始学习率设置为 0.000 01，设置自适应的学习率衰减系数为 0.5，即随着训练次数增加学习率逐渐衰减。网络整体损失是两条流的损失之和，采用 Adagrad 优化器进行网络优化，寻找损失之和的全局最优点。将训练过程保存在指定文件中，并实现训练过程可视化，两种融合方式的训练过程如图 8 所示。

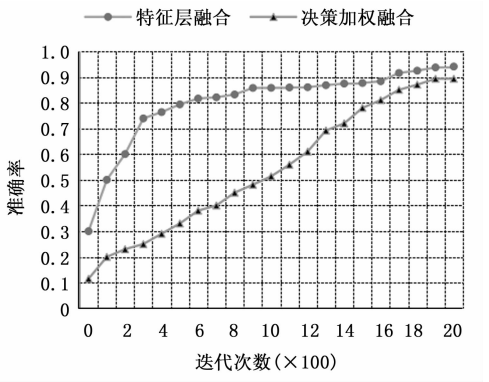

图 8 特征层融合与决策层融合训练过程

本实验的 max _ to _ keep 为 5，即每次训练保存最近的 5 个模型，输入测试集依次对每个模型进行测试。每个模型测试 10 次并记录每次的正确率和损失，取 10 次的平均值作为最终的结果，NPUAction 数据集在两种融合方式的网络中平均测试准确率如表 4 所示。然后对每一类行为分别进行测试，得出 NPUAction 数据集在两种融合方式下按行为类别的识别准确率比较图，如图 6 所示。

表 4 两种融合方式在 NPUAction 数据集上的结果

网络模型	NPUAction 平均测试准确率/%
决策层融合网络	89.29
特征层融合网络	94.05

由实验结果可以得到，双流融合中准确率较低的决策加权融合网络比单流网络中表现最好的三通道数据 88.75%的准确率高出 0.54%，说明了双流网络能有效融合 RGBD 数据中两种模态数据的优势，提高了行为识别的性能。特征拼接融合方式的准确率比决策加权融合方式高出 4.76 个百分点，取得了 94.05%的准确率，达到了目前主流行为识别算法的水平。

4 结束语

本文首先在原始 3D 卷积网络中结合注意力机制得到 AM3D 网络，注意力机制对不同区域赋予不同的权重，有利于提高行为识别网络的性能。提出了 TAM3D 网络，将 RGB 和 Depth 两种模态数据分别作为双流网络两个分支的输入，主干网络使用结合注意力机制的 AM3D，再将融合后的特征输入网络分类层，最终得到行为识别结果。实验结果表明，本文提出的 TAM3D 算法在公开的 RGB－D 日常行为数据集上与传统算法 iDT 相比识别准确率提高 6.25%，

与深度学习算法 TSN 相比提高 3.13%，在自建智慧客厅场景 RGB－D 数据集上达到了 94.05%的准确率，取得了较好的识别效果。

参考文献：

[1] HU J F，ZHENG W S，PAN J，et al. Deep bilinear learning for rgb－d action recognition [C] //Proceedings of the European Conference on Computer Vision (ECCV). 2018：335－351.

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

[3] 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.

[4] LIU Z，ZHANG C，TIAN Y. 3D－based deep convolutional neural network for action recognition with depth sequences [J]. Image and Vision Computing，2016，55：93－100.

[5] DU W，WANG Y，QIAO Y. Rpan：An end－to－end recurrent pose－attention network for action recognition in videos [C] //Proceedings of the IEEE International Conference on Computer Vision，2017：3725－3734.

[6] DEVANNE M，PAPADAKIS P. Recognition of activities of daily living via hierarchical long－short term memory networks [C] //2019 IEEE International Conference on Systems，Man and Cybernetics (SMC)，IEEE，2019：3318－3324.

[7] WANG H，WANG L. Modeling temporal dynamics and spatial configurations of actions using two－stream recurrent neural networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：499－508.

[8] 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.

[9] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [C] //Advances in Neural Information Processing Systems，2017：5998－6008.

[10] 回佳菡，施立楠，张　朋，等. 人脑视觉意识的神经机制 [J]. 生物化学与生物物理进展，2016，43 (4)：297－307.

[11] MNIH V，HEESS N，GRAVES A. Recurrent models of visual attention [C] //Advances in Neural Information Processing Systems，2014：2204－2212.

[12] XU K，BA J，KIROS R，et al. Show，attend and tell：neural image caption generation with visual attention [C] //International Conference on Machine Learning，2015：2048－2057.

[13] 王　婷，刘光辉，张钰敏，等. 多模态特征融合的长视频行为识别方法 [J]. 计算机测量与控制，2021，29 (11)：65－170.

[14] PEI W，DIBEKLIOĞLU H，BALTRUŠAITIS T，et al. Attended end－to－end architecture for age estimation from facial expression videos [J]. IEEE Transactions on Image Processing，2019，29：1972－1984.

[15] 徐胜军，欧阳朴衍，郭学源，TAHA MUTHAR KHAN. 基于多尺度特征融合模型的遥感图像建筑物分割 [J]. 计算机测量与控制，2020，28 (7)：214－219.

[16] 王　炎，连晓峰，叶　璐. 基于特征融合的多尺度窗口产品外观检测方法 [J]. 计算机测量与控制，2017，25 (12)：39－42.

[17] 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 (CVPR)，2016：1933－1941

[18] SUNG J，PONCE C，SELMAN B，et al. Human activity detection from RGBD images [C] // Proceedings of the 16th AAAI Conference on Plan，Activity，and Intent Recognition. AAAI Press，2011：47－55.

[19] WANG H，SCHMID C. Action recognition with improved trajectories [C] //Proceedings of the IEEE International Conference on Computer Vision，2013：3551－3558.

[20] 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.