基于骨骼及表观特征融合的动作识别方法

王洪雁[1,2]，袁海[2]
（1. 浙江理工大学信息学院，浙江 杭州 310018；2. 大连大学信息工程学院，辽宁 大连 116622）

摘 要：针对传统动作识别算法不易区分相似动作的问题，提出一种基于深度关节与手工表观特征融合的动作识别方法。首先将关节空域位置及约束输入具有时空注意力机制的长短期记忆（LSTM）模型中，获取时空加权且高可分的深度关节特征；然后引入热图定位关键帧及关节，手工提取关键关节周围表观特征以作为深度关节特征有效补充；最后基于双流网络逐帧融合表观特征及深度骨骼特征来实现相似动作有效判别。仿真结果表明，与主流方法相比，所提方法能有效区分相似动作，进而显著提升动作准确率。

关键词：动作识别；长短期记忆；时空注意力机制；骨骼关节；表观特征
中图分类号：TP391
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2022020

Action recognition method based on fusion of skeleton and apparent features

WANG Hongyan[1,2], YUAN Hai[2]
1. School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
2. College of Information Engineering, Dalian University, Dalian 116622, China

Abstract: Focusing on the issue that traditional skeletal feature-based action recognition algorithms were not easy to distinguish similar actions, an action recognition method based on the fusion of deep joints and manual apparent features was considered. The joint spatial position and constraints was firstly input into the long short-term memory (LSTM) model equipped with spatio-temporal attention mechanism to acquire spatio-temporal weighted and highly separable deep joint features. After that, heat maps were introduced to locate the key frames and joints, and manually extract the apparent features around the key joints that could be considered as an effective complement to the deep joint features. Finally, the apparent features and the deep skeleton features could be fused frame by frame to achieve effectively discriminating similar actions. Simulation results show that, compared with the state-of-the-art action recognition methods, the proposed method can distinguish similar actions effectively and then the accuracy of action recognition is promoted rather obviously.

Keywords: action recognition, LSTM, spatio-temporal attention mechanism, skeleton joint, apparent feature

0 引言

作为机器视觉领域的研究热点，人体动作识别在智能监控、人机交互、自动驾驶等领域发挥重要作用[1]。基于表观序列的传统识别模型通过获取颜色纹理等来识别动作，此类方法易受光照、尺度、背景变化等因素影响，且由于深度信息缺失，因此识别性能较差[2]。针对此问题，Liu 等[3-4]提出基于

收稿日期：2021−09−27；修回日期：2021−12−13
基金项目：国家自然科学基金资助项目（No.61301258, No.61271379, No.61871164）；浙江省自然科学基金重点资助项目（No.LZ21F010002）；中国博士后科学基金资助项目（No.2016M590218）
Foundation Items: The National Natural Science Foundation of China (No.61301258, No.61271379, No.61871164), The Key Projects of Natural Science Foundation of Zhejiang Province (No.LZ21F010002), China Postdoctoral Science Foundation (No.2016M590218)

深度图的识别方法，深度图所含深度信息对光照、背景变化具有较好稳健性，识别性能较好，但其信息冗余导致计算复杂，从而限制了此类方法的实际应用。

为解决上述问题，Shotton 等[5-7]提出低冗余高可分关节信息表示可显著提升动作识别性能。Vemulapalli 等[8]利用 3D 关节坐标分析运动模式识别动作，所采用的运动信息提取方法简单高效，然而该方法忽略了关节间空域关系从而有限提升准确率。针对此问题，Ahmed 等[9]采用相对距离及角度编码关节改善准确率，然而其仅依赖手工特征的识别结果难以令人满意。随着人工智能快速发展，深度学习模型利用非线性神经网络抽取深层次动作特征提升准确率[10]。其中，基于卷积神经网络（CNN, convolutional neural network）优良的空域特征提取能力，Banerjee 等[11]将骨骼序列编码为伪图像，并基于 CNN 抽取其深度特征以改进识别效果，然而所得编码图像缺失时域信息，导致准确率提升有限。针对此问题，具有良好时间建模能力的循环神经网络（RNN, recurrent neural network）可以较高准确率识别动作，然而 RNN 所固有的梯度弥散缺陷使其难以学习较长历史信息[12]。基于此，长短期记忆（LSTM, long short-term memory）模型重构 RNN 时序信息传递结构以获得优异的长时依赖关系刻画能力，可有效应用于动作识别[13-15]。Kwak 等[16]将关节时序编码为图像序列，利用 LSTM 模型抽取其时域特征改善识别性能。然而，上述基于深度网络的识别方法逐帧处理各幅图像，缺乏对关键图像及部位的挖掘，而动作序列通常存在较大信息冗余，使相关方法实时性较差且所获取高可分信息匮乏，导致准确率提升有限。基于此，Song 等[17]提出基于时空注意力机制的 LSTM（STA-LSTM, spatio-temporal attention LSTM）模型，采用时空注意力机制抽取骨骼特征，并基于重要性赋予关节相应权重以增强关键图像及部位影响，从而提升动作准确率。然而，该方法仅考虑关节坐标而忽略空域拓扑信息，准确率改善有限。此外，上述基于 3D 骨骼的相关算法仅考虑骨骼深度信息，忽略了有效表达动作的外观特征。

针对上述问题，本文提出基于骨骼关节及表观特征融合的双流网络动作识别方法。所提方法首先基于关节空间拓扑构建空域约束；其次将所得空域约束及关节坐标转化为伪图像，并输入具有时空注意力机制的 LSTM 模型以降低信息冗余，同时增强关键图像及关节的重要性提升关节深度特征表达有效性；再次基于时空注意力机制引入热图，定位图像重要关节点以提取其周围颜色纹理等外观特征；最后基于双流网络逐帧融合表观及关节深度特征序列以实现复杂场景下人体动作有效识别。

基于以上所述，本文贡献可简述如下。

1) 利用所构建关节相对距离与高相关度关节对等空域约束有效补充骨骼时空动态信息，并将其转化为伪图像。

2) 构建具有时空注意力机制的 LSTM 模型，采用时序权重差值法去除相似帧，基于热图定位序列关键帧及关节，并以所得关键关节作为表观特征提取区域。

3) 基于双流网络逐帧融合手工表观特征及 LSTM 所得深度骨骼特征序列以有效识别相似动作。

1　基于关节及外观特征融合的识别模型

所提动作识别模型主要包含如下 4 个部分：首先，构建关节空间约束，即关节相对距离与高相关度关节对；其次，构建具有时空注意力机制的 LSTM 模型；再次，基于热图定位重要关节并抽取附近颜色纹理等外观特征；最后，基于双流网络逐帧融合骨骼序列所得关节特征及表观序列所得外观特征以提升动作准确率。模型如图 1 所示。

1.1　关节空间约束

1.1.1　关节坐标

关节信息可有效表征人体姿态，从而可作为动作高可分表达，通过将动态关节信息输入深度网络以获取关节序列的深度有效特征，从而提升动作准确率。人体结构可分为左臂、右臂、躯干、左腿、右腿 5 个部分，对于全部关节点 K（本文中 $K=25$），$\boldsymbol{X}_{t,k}=(x_{t,k},y_{t,k},z_{t,k})$ 表示第 t（$t=1,2,\cdots,T$）帧内关节 k 的坐标，则所有关节坐标可表示为 $\boldsymbol{X}_t=(\boldsymbol{X}_{t,1},\cdots,\boldsymbol{X}_{t,K})$，其中 T 为序列帧数。

1.1.2　相对距离约束

众所周知，无论是静止还是运动状态，关节间始终具有特定范围内距离关系，因此关节相对距离可有效表示人体局部感兴趣区域，并且对视角及光照变化具有较强稳健性。此外，运动过程中髋关节 $\boldsymbol{X}_{t,1}=(x_1,y_1,z_1)$ 变化幅度较小，其余关节均围绕髋关节做定向圆周运动，因此，可将其取为坐标中心。由此，髋关节与其他关节之间的欧氏距离可表示为

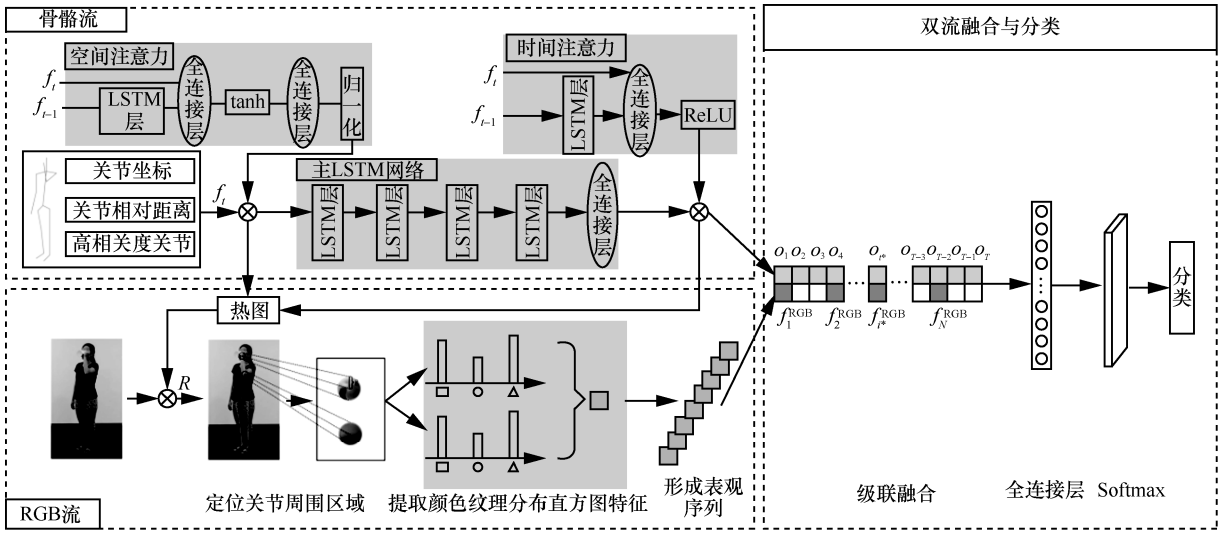


图 1 基于深度关节特征及手工外观特征融合的动作识别模型

$$d_{t,j\_1}=\sqrt{(x_j-x_1)^2+(y_j-y_1)^2+(z_j-z_1)^2} \quad (1)$$

其中，$j=2,3,\cdots,K$。

为避免个体间身高差异，归一化 d_{t,j_1} 可得如下相对距离

$$\boldsymbol{l}_{t,j\_1}=\frac{\left|d_{t,j\_1}\right|}{\left|d_{t,21\_1}\right|} \quad (2)$$

其中，$d_{t,21\_1}$ 为锁骨及髋关节的距离。由此，动作序列中第 t 帧内关节相对距离可表示为

$$\boldsymbol{B}_t=[\boldsymbol{l}_{t,2\_1},\boldsymbol{l}_{t,3\_1},\cdots,\boldsymbol{l}_{t,K\_1}] \quad (3)$$

1.1.3 高相关度关节对约束

人体骨骼中任意关节间皆存在一定数量骨骼边，某关节的运动将导致相邻关节同步运动，两关节间相连边越少，表明关节间距离较近，协作关系更密切、相关度更高。基于此观察，本文只选取相关度较高的一、二（即只有一或两条边相连关节对）级相关信息构建关节空域相关约束以降低计算复杂度，其中关节相对位置为

$$\boldsymbol{C}_{t,i\_j}=\boldsymbol{X}_{t,i}-\boldsymbol{X}_{t,j} \quad (4)$$

其中，$\boldsymbol{C}_{t,i\_j}$ 表示第 t 帧内第 j 个关节相对第 i 个关节的坐标，即二者空域拓扑信息。

综上所述，一、二级相关信息分别为

$$\boldsymbol{R}_1=\left[\boldsymbol{C}_{h\_k},\boldsymbol{C}_{m\_n},\cdots,\boldsymbol{C}_{o\_p}\right],\boldsymbol{R}_2=\left[\boldsymbol{C}_{q\_r},\boldsymbol{C}_{u\_v},\cdots,\boldsymbol{C}_{x\_y}\right] \quad (5)$$

其中，h_k、m_n、o_p 等表示仅由一条边相连的关节对，q_r、u_v、x_y 等表示由两条边相连的关节对。

综上所述，有效表征某动作的关节序列时空信息可表示为

$$\boldsymbol{f}_t=[\boldsymbol{X}_t;\boldsymbol{B}_t;\boldsymbol{R}_1;\boldsymbol{R}_2] \quad (6)$$

通常认为，整个动作期间可有效表达动作的图像帧及关节更具重要性[18]，以序列“跳跃”为例，相较于直立帧及躯干，跳跃帧及四肢更具指标意义。基于此，本节提出如图 2 所示的基于时空注意力的 LSTM 模型以加权各帧及部位从而体现其重要性。

1.2 具有空间约束的时空注意力 LSTM 模型

1.2.1 空间注意力

如上所述，视频帧及各关节对动作识别影响不同，基于此事实，本节基于空间注意力机制加权各关节以反映其重要程度从而增强动作可区分度。设时刻 t 所有关节权重为 $\boldsymbol{\alpha}_t=(\alpha_{t,1},\cdots,\alpha_{t,l})$，$l$ 为输入特征 $\boldsymbol{f}_t$ 维数，对应得分 $\boldsymbol{s}_t=(s_{t,1},\cdots,s_{t,l})$ 可表示为

$$\boldsymbol{s}_t=\tanh(\boldsymbol{w}_f\boldsymbol{f}_t+\boldsymbol{w}_h\boldsymbol{h}_{t-1}+\boldsymbol{b}) \quad (7)$$

其中，为避免前向传播数值上溢问题采用 tanh 激活函数，$\boldsymbol{w}_f$、$\boldsymbol{w}_h$ 分别为输入数据 $\boldsymbol{f}_t$ 及上层 LSTM 隐藏变量 $\boldsymbol{h}_{t-1}$ 的加权矢量，$\boldsymbol{b}$ 为偏差矢量。

基于上述关节得分，经由 Softmax 计算，可得如下可有效表征关节空域重要性的权值

$$\alpha_{t,i}=\frac{\mathrm{e}^{s_{t,i}}}{\sum_{j=1}^{l}\mathrm{e}^{s_{t,j}}} \quad (8)$$

由此可得如下输入主 LSTM 模型的空域加权特征

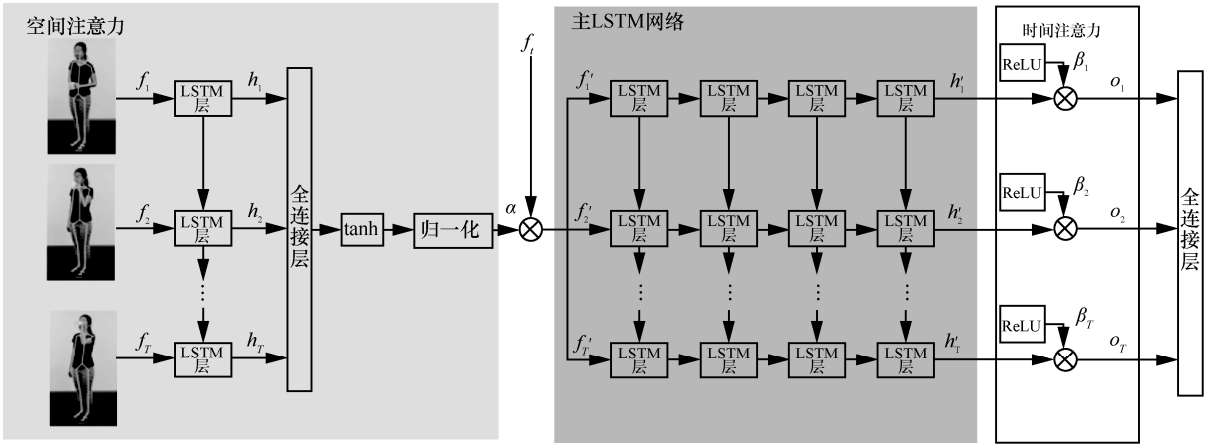


图 2　基于时空注意力的 LSTM 模型

$$\boldsymbol{f}_t' = \boldsymbol{\alpha}_t \odot \boldsymbol{f}_t \tag{9}$$

其中，$\odot$ 为 Hadamard 积，表示矢量相应元素相乘。

1.2.2　时间注意力

动作识别过程中视频序列存在大量冗余帧，针对此问题，本节利用时间注意力机制加权序列以突出关键帧同时降低信息冗余度从而提升动作准确率。各帧权重 β_t 可表示为

$$\beta_t = \mathrm{ReLU}(\tilde{\boldsymbol{w}}_f \boldsymbol{f}_t + \tilde{\boldsymbol{w}}_h \tilde{\boldsymbol{h}}_{t-1} + \tilde{\boldsymbol{b}}) \tag{10}$$

其中，ReLU 为非线性激活函数，$\tilde{\boldsymbol{h}}_{t-1}$ 为上一帧隐藏变量，$\tilde{\boldsymbol{w}}_f$ 和 $\tilde{\boldsymbol{w}}_h$ 为待学习参数，$\tilde{\boldsymbol{b}}$ 为偏差向量。

基于以上所述，如图 2 所示，$\boldsymbol{f}_t'$ 经由主 LSTM 模型的输出为 $\boldsymbol{h}_t'$，对其加以时间注意力机制，所得输出序列特征可表示为 $\boldsymbol{o}_t = \beta_t \boldsymbol{h}_t'$。

1.3　手工表观特征构造

动作识别中颜色纹理特征可直观反映姿态变化，由此可将包含丰富颜色及纹理信息的表观序列作为基于骨骼信息动作识别的有效补充。若对整幅图像提取外观特征，则难以直观反映动作细微差异。基于此，本节利用热图定位关键帧及关节（如图 3 所示），并在其附近半径为 R 的圆形区域提取颜色纹理直方图，作为关节深度特征的有效补充。

由于关键帧通常处于稳态且相邻帧差异较小，因此应避免提取大量相似帧以降低计算复杂度同时改善准确率。本节以各帧时间注意力权重差值为区分准则来划分相似帧片段，并提取片段中权重最大帧来表征相似帧片段。注意到，相邻帧越相似、权重值越相近，则其差值越小。基于此，权重为 β_i（$1 \leqslant i \leqslant T$）的序列帧 i 与参考帧 β_{i^*}（参考帧为各片段首帧，$1 \leqslant i^* \leqslant N$）之间的权重差值为 β_c，即

$$\beta_c = \left| \beta_i - \beta_{i^*} \right| \tag{11}$$

基于此，令 δ 为相似帧权重差值阈值，当 $\beta_c < \delta$ 时，表明后续帧和当前参考帧类似；当 $\beta_c \geqslant \delta$ 时，帧 i^* 为新参考帧，最终提取所有参考帧 N 构成关键帧。

需要注意的是，关键帧内不同权重关节可影响相似动作判别，由各关节权重所得热图则表征了重要关节运动趋势，如图 3 所示相似动作中具有代表性的三帧，其手部周围区域体现相似动作细微差异。基于此，通过提取手部颜色纹理特征，并加以关节点权重 $\boldsymbol{\alpha}_{i^*} = (\alpha_{i^*,1}, \cdots, \alpha_{i^*,K})$ 以增强外观信息，从而可有效获取手物信息以作为关节特征的有力补充。

1.3.1　LBP 纹理特征

由于局部二值模式（LBP, local binary pattern）具有灰度不变及旋转不变性[19]，光照变化稳健性较好，因而在图像识别领域得到广泛应用[20-21]。基于此，重要关节附近纹理可基于 LBP 表达。设 n_c 为中心点灰度值，$n_0 \sim n_7$ 为邻域点灰度值。以 n_c 为阈值依次比较邻域像素点，若像素灰度值大于阈值将该点标记为 1，否则为 0。将结果采用顺时针构成二进制序列，作为该点 LBP 值，计算式为[18]

$$\mathrm{LBP}(x,y) = \sum_{i=0}^{N-1} s(n_i - n_c) 2^i \tag{12}$$

$$s(x) = \begin{cases} 1, & x > 0 \\ 0, & \text{其他} \end{cases} \tag{13}$$

其中，$\boldsymbol{M}_{i^*} = (\boldsymbol{M}_{i^*,1}, \cdots, \boldsymbol{M}_{i^*,K})$ 为第 i^*（$i^* = 1,2,\cdots,N$）关键帧圆形区域对应纹理直方图向量。

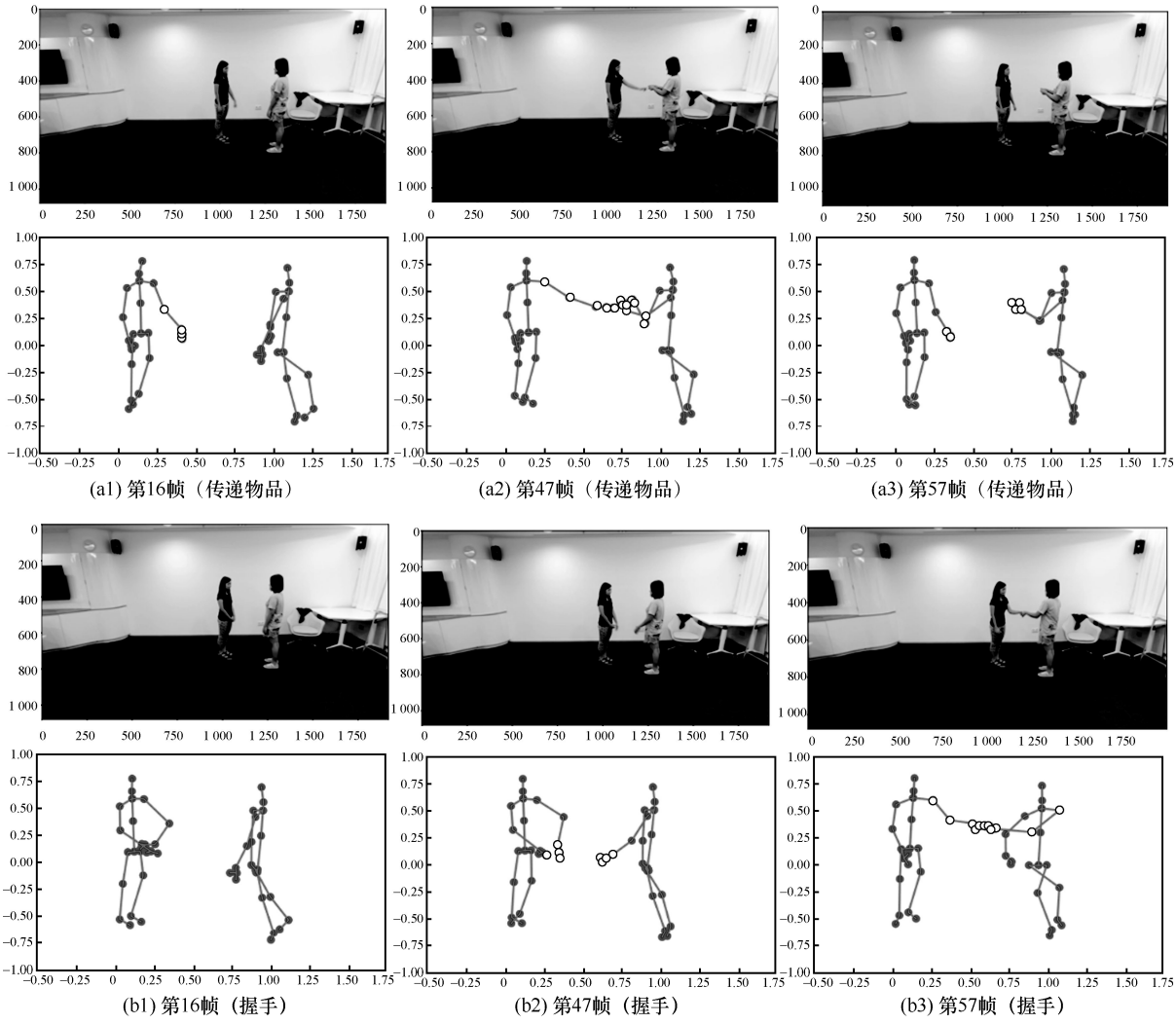


图 3　基于热点定位重要关节

1.3.2　HSV 颜色直方图

颜色直方图可有效描述各色彩比例，HSV 颜色模型将亮度色度分离，因而不易受光照变化等因素干扰[22]。基于此，本节基于 HSV 颜色空间模型构建颜色特征。HSV 空间中色调 H 较饱和度 S 及亮度 V 敏感，故赋予 H 通道更多量化级别。此外，量化间隔越大则信息损失越多；间隔越小则信息损失越少，同时数据量显著增加，进而导致计算复杂度上升。由此，本节基于文献[23]构造如下量化等级

$$\boldsymbol{H}=\begin{cases}0, h\in[316,360]\cup[0,20]; & 1, h\in[21,40]\\ 2, h\in[41,75]; & 3, h\in[76,155]\\ 4, h\in[156,190]; & 5, h\in[191,270]\\ 6, h\in[271,295]; & 7, h\in[296,315]\end{cases} \tag{14}$$

$$\boldsymbol{S}=\{0, s\in[0,0.2];1, s\in[0.2,0.7];2, s\in[0.7,1]\} \tag{15}$$

$$\boldsymbol{V}=\{0, v\in[0,0.2];1, v\in[0.2,0.7];2, v\in[0.7,1]\} \tag{16}$$

将上述非均匀量化 HSV 合成如下矢量 $\boldsymbol{G}$

$$\boldsymbol{G}=HQ_sQ_v+SQ_v+V \tag{17}$$

其中，Q_s、Q_v 分别为 S、V 分量量化级数。

由式(14)～式(16)可知，HSV 分别量化为 8、3 和 3 级，则 $Q_s=3$，$Q_v=3$。同时 HSV 分别取最大值 7、2 和 2，则 $\boldsymbol{G}$ 取值范围为[0, 71]。基于此，可将 HSV 空间表述为包含颜色级别为 72 的特征向量，统计该颜色级别频率以获得 HSV 颜色直方图，则 $\boldsymbol{G}_{i^*}=(\boldsymbol{G}_{i^*,1},\cdots,\boldsymbol{G}_{i^*,K})$ 为各子块对应直方图向量。

热图所指示关键关节周围提取颜色纹理分布直方图，为保证局部区域性质，可先拼接单个圆形区域，再将表观序列圆形区域乘以对应关节点权重依次连接可获得参考帧颜色纹理特征（如图 4 所示）

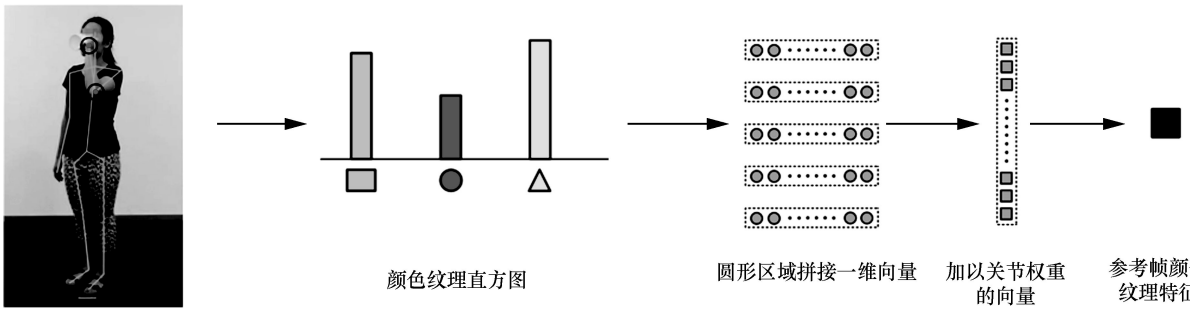


图 4　颜色纹理直方图融合

$$\boldsymbol{f}_{i^*}^{\mathrm{RGB}} = \mathrm{Concat}(\alpha_{i^*}(\boldsymbol{M}_{i^*}, \boldsymbol{G}_{i^*})) \tag{18}$$

1.4　基于深度关节与表观特征双流融合

综上所述，通过具有空间约束的时空注意力机制 LSTM 模型（STA-SC-LSTM），提取运动变化关键关节特征，基于热图定位表观关键帧及重要关节以手动提取重要关节周围颜色纹理等表观细节信息，所提动作识别模型基于双流网络融合所得深度关节及表观特征。

根据表观及深度特征特殊对应关系，本节采用更利于提升准确率的逐帧融合再序列融合方法，以突出局部重要部位互补性。根据上述权重差值 β_c 判定各段相似帧的参考帧 β_{i^*}，同时记录各段相似帧数量 φ_{i^*}（$1 \leqslant i^* \leqslant N$），则表观序列参考帧位置 i^* 对应由 LSTM 模型提取关节深度特征序列位置 $t^* = \sum_{i^*=1}^{N} \varphi_{i^*-1}$（当 i^* 为 1 时，$\varphi_0 = 1$）。基于此，参考帧表观特征 $\boldsymbol{f}_{i^*}^{\mathrm{RGB}}$ 与对应深度关节特征 $\boldsymbol{o}_{t^*}$ 以权重占比 λ_2 与 λ_1 融合。其中，$\lambda_1+\lambda_2=1$（二者可经由实验确定，具体参见实验部分），对应帧融合特征可表示为

$$\boldsymbol{x}_{i^*} = \mathrm{Concat}(\lambda_1 \boldsymbol{o}_{t^*}, \lambda_2 \boldsymbol{f}_{i^*}^{\mathrm{RGB}}) \tag{19}$$

同时，无参考帧对应的深度特征补 0 以降低系统复杂性。最后序列融合特征 $\boldsymbol{x}_i^{\mathrm{fusion}}$（其中，$i = 1, \cdots, C$，$C$ 表示动作类别数）映射至全连接层并基于 Softmax 函数识别动作

$$p\left(\boldsymbol{x}_i^{\mathrm{fusion}}\right) = \frac{\exp(\boldsymbol{x}_i^{\mathrm{fusion}})}{\sum_{j=1}^{C} \exp(\boldsymbol{x}_j^{\mathrm{fusion}})} \tag{20}$$

为提升训练效果，构造如下正则化损失函数

$$L = -\sum_{i=1}^{C} y_i \log \hat{y}_i + \lambda \| \boldsymbol{W} \|_2 \tag{21}$$

其中，第一项基于交叉熵 $\boldsymbol{y} = (y_1, \cdots, y_C)^{\mathrm{T}}$ 为真实动作，$\hat{y}_i = p(\boldsymbol{x}_i^{\mathrm{fusion}})$ 为第 i 类动作预测概率；第二项为模型参数正则化约束以抑制过拟合，λ 为损失函数平衡因子，$\boldsymbol{W}$ 为模型参数。

2　实验结果及分析

基于 NTU RGB-D、Northwestern-UCLA、SBU Interaction Dataset 这 3 个公开动作识别数据集，本节通过与基于手工特征、CNN、RNN 及 LSTM 等模型的动作识别方法在视角变化、主体多样化及同类动作多样化等方面对比，验证所提方法有效性。

2.1　实验环境

本节实验基于 TensorFlow 深度学习框架，处理器 Intel Core(TM) i7-7700，主频 3.60 GHz，32 GB 内存，NVIDIA GeForce GTX 1070。选取 4 层 LSTM 作为主网络，时空注意力分别基于单个 LSTM，每层神经元个数均为 128，表观特征提取半径为 5 像素点，初始学习率为 0.002，每经过 30 次训练学习率缩小至 10%，采用动量为 0.8 的随机梯度下降法作为优化函数 Adam，平衡因子 $\lambda = 10^{-5}$，批处理大小为 64，Dropout=0.45 以防止过拟合。

2.2　NTU RGB-D 数据集

NTU RGB-D 数据集为目前包含受测对象和行为类别数目最大的 RGB-D 行为数据集[24]。该数据集由 40 位受测对象通过 3 台 Kinect V2 摄像机从−45°、0°、45°这 3 个不同角度采集 60 类动作，56 880 个视频片段与三维骨骼数据序列。其中包括个体日常动作（如跌倒、呕吐、鼓掌等）、人物交互（如梳头、撕纸、踢东西等）、双人交互（如推、拍后背、手指对方等），以及诸如喝水与刷牙、阅读与写作、握手与传递物品等具有细微差别的动作。

交叉主体（cross subject）实验将 40 类受测对象分为训练及测试集[24]，训练集编号为 1、2、4、5、

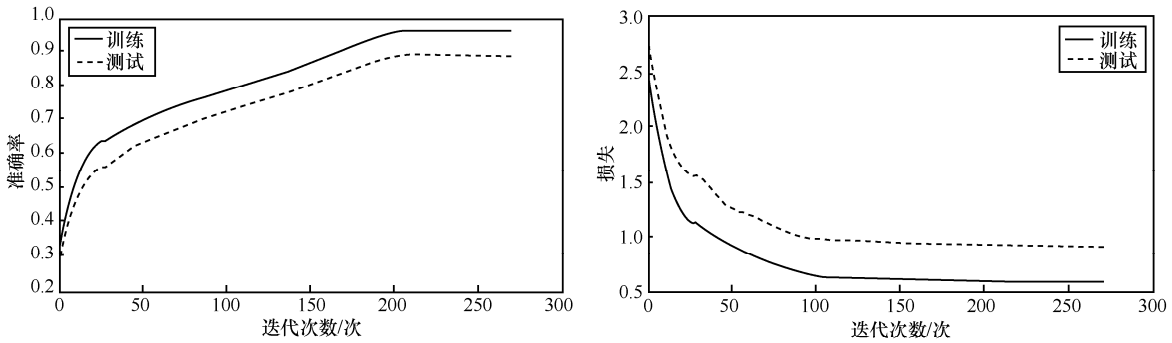
8、9、13、14、15、16、17、18、19、25、27、28、31、34、35、38，其余为测试集，训练集和测试集分别为 40 320 和 16 560 个样本；交叉视图（cross view）实验选取第一台摄像机采集样本为测试集，其余为训练集，训练集和测试集分别为 37 920 和 18 960 个样本。

本节实验交叉主体与交叉视图迭代训练中训练集与测试集对应的准确率与损失曲线如图 5 所示。由图 5 可知，模型准确率随着训练次数增加而增加，迭代至 220 次时准确率趋于稳定且损失值收敛。此外，基于 NTU RGB-D 数据集可得交叉主体及交叉视角准确率分别为 88.73%和 90.01%，其识别结果可由图 6 所示的混淆矩阵表征。

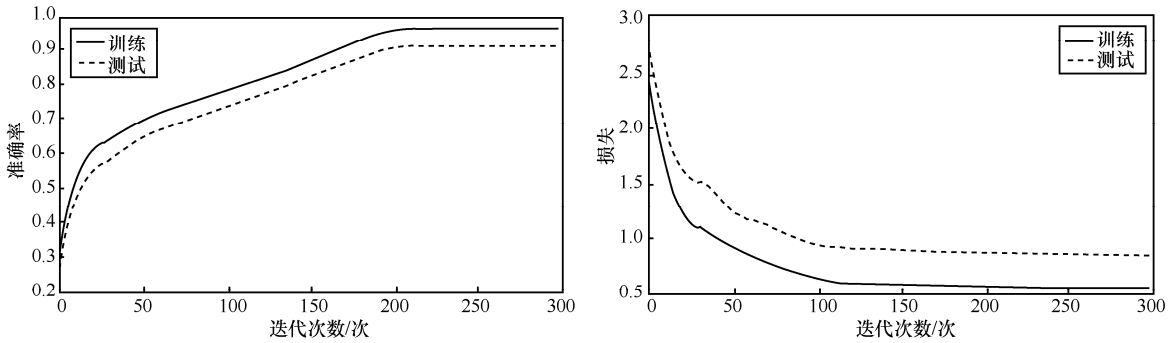
图 6 中各列及各行分别为所提方法预测结果及对应真实类别，主对角线元素表示该动作准确率，其余为识别错误率。由图 6 可知，交互相似动作，即喝水、刷牙与打电话，阅读、写作、键盘打字与玩手机的交叉主体及交叉视角准确率分别不低于 84%和 86%；双人交互相似行为，即握手和传递物品的交叉主体及交叉视角准确率分别不低于 80%和 88%。此外，其他动作交叉主体及交叉视角准确率分别为 85%～92%和 87%～94%。由此可知，主题多样化及视角变化等复杂场景下所提方法具有较高准确率。

基于 NTU RGB-D 数据集，所提方法及主流方法所得交叉主体及交叉视角准确率如表 1 所示。

由表 1 可知，基于可变参数关联骨架的 LARP（lie group action recognition point）[8]与基于 3D 几何关系的 Dynamic skeletons[25]没有考虑深度时空信息，因而准确率不高；Multi temporal 3D CNN 将关节映射至 3D 空间并通过 3D CNN 提取深度特征，从而可有效提升准确率至 66.85%、72.58%[26]，然而其没有考虑骨骼识别时域信息；ST-LSTM+Trust Gate[27]与 Two-Stream RNN[28]分别以相关关节作为双流 RNN 输入以充分利用时空信息，然而输入时序存在较大信息冗余，从而影响识别效果；基于此，STA-LSTM[17]基于时空注意力机制以识别关键帧及关节，从而将准确率提升至 73.40%、81.20%，然而该方法只考虑关节特征而忽略拓扑关系，故准确率改善有限；DS-LSTM（denoising sparse LSTM）[15]考虑帧间帧内关节链接相对运动趋势，Fuzzy fusion+CNN[11]编码关节间空间关系以提升准确率，然而二者缺乏外观特征，从而限制识别能力；所提方法将空间约束输入具有时空注意力机制的 LSTM 模型以抽取深度时空特征，并基于热图抽取表观特征为有效补充，从而提升准确率至 88.73%、90.01%，这表明复杂场景下所提方法具有较高准确率。

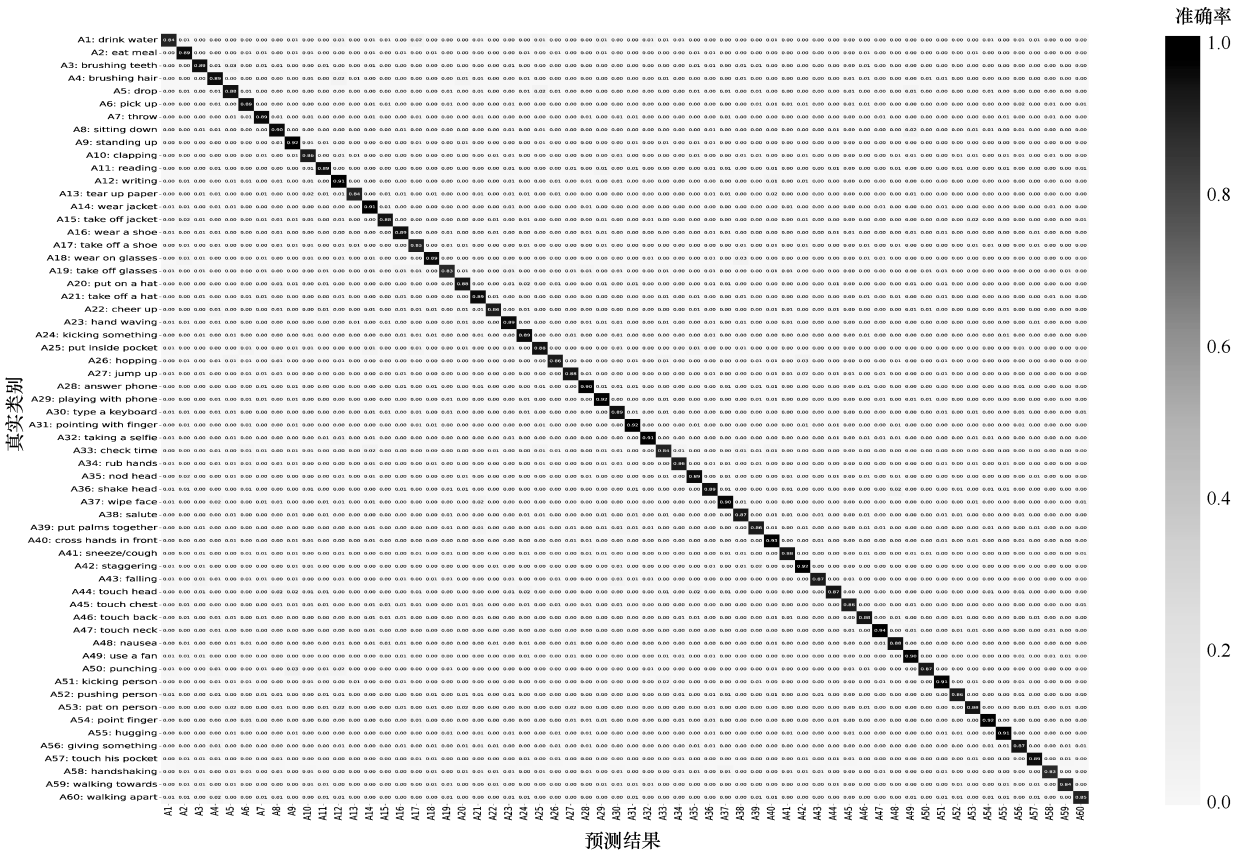


(a) 交叉主体

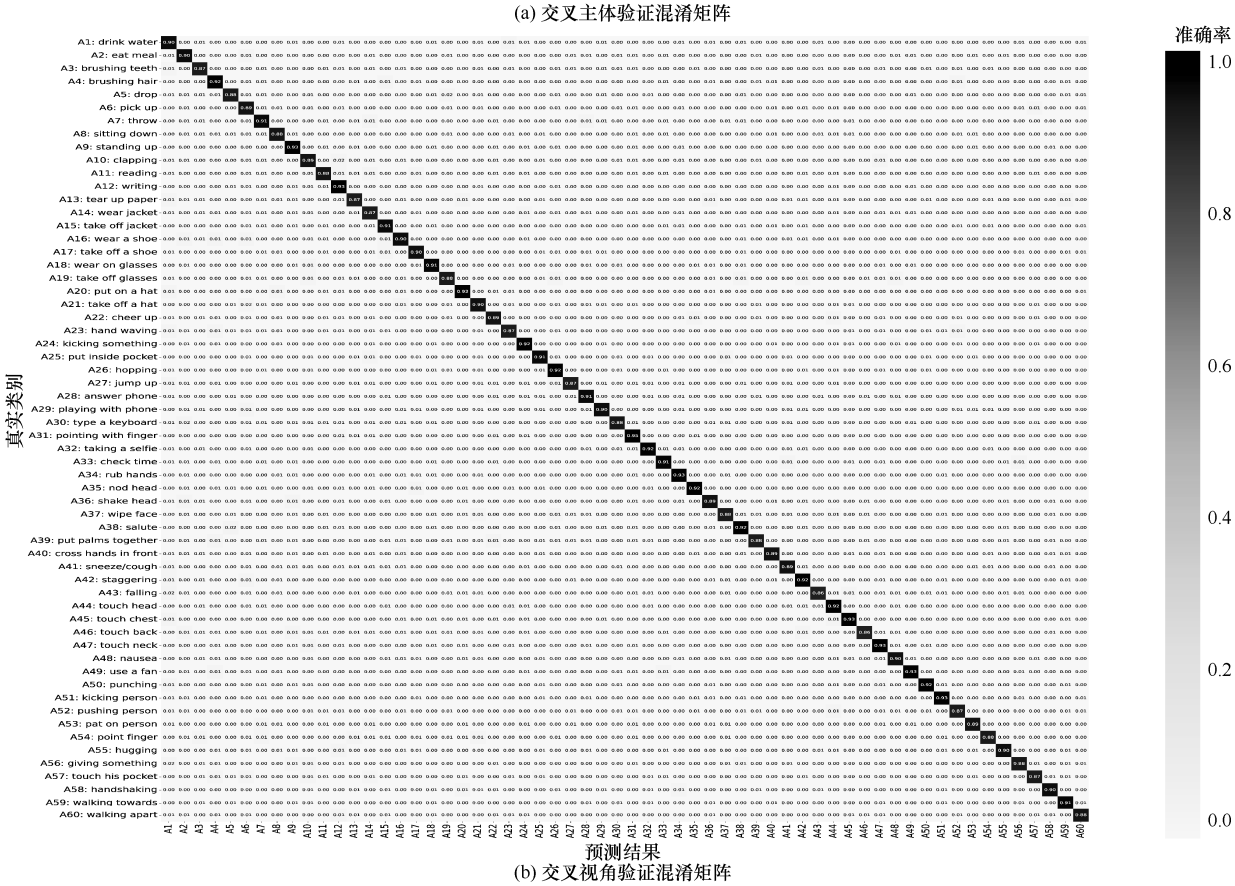


(b) 交叉视图

图 5 NTU RGB-D 数据集中训练集与测试集对应的准确率与损失曲线



(a) 交叉主体验证混淆矩阵



(b) 交叉视角验证混淆矩阵

图 6　基于 NTU RGB-D 数据集所得混淆矩阵

表 1 NTU RGB-D 数据集各方法所得交叉主体及交叉视角准确率

数据	特征	方法	cross subject	cross view
骨骼序列	手工提取	LARP	50.08%	52.76%
		Dynamic skeletons	60.23%	65.22%
	CNN	Multi temporal 3D CNN	66.85%	72.58%
	LSTM	ST-LSTM+Trust Gate	69.20%	77.70%
	RNN	Two-Stream RNN	71.30%	79.50%
	CNN	TSRJI	73.30%	80.30%
	LSTM	STA-LSTM	73.40%	81.20%
	LSTM	DS-LSTM	77.80%	87.33%
	CNN	Fuzzy fusion+CNN	84.22%	89.71%
	LSTM/手工提取	所提方法	**88.73%**	**90.01%**

2.3 Northwestern-UCLA 数据集

Northwestern-UCLA 数据集由 1 494 个序列组成，由 10 名实验者完成如下 10 类动作[30]：单手捡、双手捡、扔垃圾、行走、坐下、站起、穿、脱、扔、拿。该数据集由 3 个不同视角采集获得，前 2 个摄像机所得样本为训练数据，其余样本为测试数据。

如表 2 所示，基于骨骼特征手工提取的 HOJ3D（histograms of 3D joint）方法[31]假设骨骼垂直于地面以投影聚类判别动作，忽略骨骼空域关系，从而导致准确率较低；LARP[8]则基于可变参数关联骨架表征动作，因而性能优于 HOJ3D，但是其忽略骨骼动态信息；HBRNN-L（hierarchically bidirectional RNN LSTM）[32]考虑关节时域特征，从而获得78.52%的准确率，但是其缺乏外观信息难以区分相似动作；Multi-view dynamics+CNN[33]提取多视角动态图像以应对空域变化，考虑外观特征，但是其缺乏时序特征；所提方法基于具有时空注意力机制的 LSTM 模型以有效表征重要关节动态信息，并基于热图抽取颜色纹理信息，从而获得动作高可分表达，进而将准确率提升至 85.73%，分别比 HBRNN-L 和 Multi-view dynamics+CNN 提升 7.21%和 1.53%，这表明不同视角及主题多样化条件下所提方法具有较高识别能力。

表 2 Northwestern-UCLA 数据集实验结果

数据	特征	方法	准确率
骨骼序列	手工提取	HOJ3D	54.50%
		LARP	74.20%
	RNN/LSTM	HBRNN-L	78.52%
	CNN	Multi-view dynamics+CNN	84.20%
	LSTM/手工提取	所提方法	**85.73%**

2.4 SBU Interaction 数据集

SBU Interaction 数据集包含如下 8 类交互动作[34]：靠近、远离、踢、推、握手、拥抱、递书本、拳击，共分为 5 个交叉集，选取其中 4 个作为训练集，其余为测试集，对各交叉集验证结果取平均值作为最终准确率。

所提方法及对比方法所得准确率如表 3 所示。由表 3 可知，所提方法准确率可达 95.46%，分别比 STA-LSTM[17]、ST-LSTM+Trust Gate[27]、Two-Stream RNN[28]提升 3.96%、2.16%、0.66%，这表明小样本数据集下所提方法准确率较高。

2.5 消融实验

为进一步验证所提方法有效性，基于上述数据集研究所提方法中具有空间约束时空注意力 LSTM 模型及特征融合模块对准确率影响，所得结果如表 4 所示。由表 4 可知，相较于仅基于时空注意力的 STA-LSTM 模型，STA-SC-LSTM 所得准确率分别提升 2.43%、1.52%、0.83%，表明所构造空域约束条件可提升动作识别能力；相较于仅基于关节时序特征的 STA-SC-LSTM，双流融合所得准确率分别提升 12.90%、7.29%、8.15%及 3.13%，表明表观特征可作为骨骼深度特征的有效补充以弥补基于关节时空特征的相关模型对相似动作较低区分度的缺陷。

表 3 SBU Interaction Dataset 数据集实验结果

方法	准确率
Joint Feature	86.90%
Co-occurrence Feature	90.40%
STA-LSTM	91.50%
ST-LSTM+Trust Gate	93.30%
Two-Stream RNN	94.80%
所提方法	**95.46%**

表 4 不同模型实验结果

数据集	STA-LSTM	STA-SC-LSTM	双流融合
NTU（cross subject）	73.40%	75.83%	**88.73%**
NTU（cross view）	81.20%	82.72%	**90.01%**
Northwestern-UCLA	—	77.58%	**85.73%**
SBU	91.50%	92.33%	**95.46%**

2.6 双流融合权重设置

融合深度关节及手工表观特征可有效提升相似动作判别性能，然而融合权重难以由理论确定。由此，本节基于上述数据集，通过实验确定融合权值。具体地，(λ_1,λ_2) 可依次设为（0.4, 0.6）、（0.5, 0.5）、

（0.6, 0.4）和（0.7, 0.3）。由表 5 可知，权重由（0.4, 0.6）变化至（0.5, 0.5），即关节特征权重占比增加则准确率提升，表明识别结果主要依赖于关节特征。当权重由（0.6, 0.4）变化至（0.7, 0.3），准确率降低，表明外观特征缺乏，从而影响相似动作区分。由上述分析可知，权重为（0.6, 0.4）时识别精确度最高，由此设定 $\lambda_1 = 0.6$、λ_2=0.4 为融合权重。

表 5　　不同融合权重比的实验结果

双流融合比重	NTU（cross subject）	NTU（cross view）	Northwestern-UCLA	SBU
(0.4, 0.6)	75.42%	79.93%	75.21%	93.84%
(0.5, 0.5)	84.07%	85.53%	82.29%	93.84%
(0.6, 0.4)	**88.73**%	**90.01**%	**85.73**%	**95.46**%
(0.7, 0.3)	81.34%	84.15%	81.91%	93.06%

3　结束语

本文提出基于关节序列深度时空及表观特征融合的动作识别方法。所提方法首先构建关节空域拓扑约束以增强关节特征表达有效性，其次构造具有时空注意力的 LSTM 以定位高可分重要帧及关节，再次基于热图提取关键关节周围颜色纹理表观特征，最后逐帧融合关节深度及外观特征以获得高可分的动作有效表达。实验结果表明，在 NTU RGB-D、Northwestern-UCLA 以及 SBU Interaction Dataset 数据集上，所提方法的准确率分别为 88.73%、90.01%，85.73%和 95.46%，明显高于现有主流识别方法，表明视角变化、噪声、主体多样化等复杂场景下所提方法的有效性。需要注意的是，由实验可知，相较于交叉主体，交叉视图准确率改善幅度较小，基于此，未来研究将着重关注多视角场景下表观及关节高可分稳健特征抽取及有效融合方法。

参考文献：

[1] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169-180.
LUO H L, WANG C J, LU F. Survey of video behavior recognition[J]. Journal on Communications, 2018, 39(6): 169-180.

[2] JIANG Y G, DAI Q, LIU W, et al. Human action recognition in unconstrained videos by explicit motion modeling[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2015, 24(11): 3781-3795.

[3] LIU M Y, LIU H. Depth Context: a new descriptor for human activity recognition by using sole depth sequences[J]. Neurocomputing, 2016, 175: 747-758.

[4] CHEN C, LIU M Y, LIU H, et al. Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition[J]. IEEE Access, 2017, 5: 22590-22604.

[5] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images[C]//Machine Learning for Computer Vision. Berlin: Springer, 2013: 119-135.

[6] HAN F, REILY B, HOFF W, et al. Space-time representation of people based on 3D skeletal data: a review[J]. Computer Vision and Image Understanding, 2017, 158: 85-105.

[7] KE Q H, BENNAMOUN M, AN S J, et al. Learning clip representations for skeleton-based 3D action recognition[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(6): 2842-2855.

[8] VEMULAPALLI R, ARRATE F, CHELLAPPA R. Human action recognition by representing 3D skeletons as points in a lie group[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 588-595.

[9] AHMED F, PAUL P P, GAVRILOVA M. Adaptive pooling of the most relevant spatio-temporal features for action recognition[C]//Proceedings of 2016 IEEE International Symposium on Multimedia. Piscataway: IEEE Press, 2016: 177-180.

[10] WANG L, HUYNH D Q, KONIUSZ P. A comparative review of recent kinect-based action recognition algorithms[J]. IEEE Transactions on Image Processing, 2020, 29: 15-28.

[11] BANERJEE A, SINGH P K, SARKAR R. Fuzzy integral-based CNN classifier fusion for 3D skeleton action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(6): 2206-2216.

[12] LE Q V, JAITLY N, HINTON G E. A simple way to initialize recurrent networks of rectified linear units[J]. arXiv Preprint, arXiv: 1504.00941, 2015.

[13] ZHANG J, BAI F S, ZHAO J F, et al. Multi-views action recognition on 3D ResNet-LSTM framework[C]//Proceedings of 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering. Piscataway: IEEE Press, 2021: 289-293.

[14] AVOLA D, CASCIO M, CINQUE L, et al. 2-D skeleton-based action recognition via two-branch stacked LSTM-RNNs[J]. IEEE Transactions on Multimedia, 2020, 22(10): 2481-2496.

[15] JIANG X H, XU K, SUN T F. Action recognition scheme based on skeleton representation with DS-LSTM network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(7): 2129-2140.

[16] KWAK I S, GUO J Z, HANTMAN A, et al. Detecting the starting frame of actions in video[C]//Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2020: 478-486.

[17] SONG S J, LAN C L, XING J L, et al. Spatio-temporal attention-based LSTM networks for 3D action recognition and detection[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society, 2018, 27(7): 3459-3471.

[18] SCHINDLER K, VAN GOOL L. Action snippets: how many frames does human action recognition require?[C]//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2008: 1-8.

[19] OJALA T, PIETIKÄINEN M, HARWOOD D. A comparative study of texture measures with classification based on featured distributions[J].

Pattern Recognition, 1996, 29(1): 51-59.

[20] PIETIKÄINEN M. Image analysis with local binary patterns[C]// Proceedings of the 14th Scandinavian Conference on Image Analysis. [S.l.:s.n.], 2005: 115-118.

[21] 梁淑芬, 刘银华, 李立琛. 基于 LBP 和深度学习的非限制条件下人脸识别算法[J]. 通信学报, 2014, 35(6): 154-160.
LIANG S F, LIU Y H, LI L C. Face recognition under unconstrained based on LBP and deep learning[J]. Journal on Communications, 2014, 35(6): 154-160.

[22] LEI L, PENG J, YANG B. Image retrieval based on HSV feature and regional Shannon entropy[J]. International Journal of Software Science and Computational Intelligence, 2012, 4(2): 64-80.

[23] YU P, ZHANG C, DU C H. Image retrievals based on color and texture features[C]//Proceedings of 2007 9th International Symposium on Signal Processing and Its Applications. Piscataway: IEEE Press, 2007: 1-4.

[24] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: a large scale dataset for 3D human activity analysis[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1010-1019.

[25] HU J F, ZHENG W S, LAI J H, et al. Jointly learning heterogeneous features for RGB-D activity recognition[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5344-5352.

[26] TU J H, LIU M Y, LIU H. Skeleton-based human action recognition using spatial temporal 3D convolutional neural networks[C]//Proceedings of 2018 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2018: 1-6.

[27] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3D human action recognition[C]//Computer Vision – ECCV 2016. Berlin: Springer, 2016: 816-833.

[28] WANG H S, WANG L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3633-3642.

[29] CAETANO C, BRÉMOND F, SCHWARTZ W R. Skeleton image representation for 3D action recognition based on tree structure and reference joints[C]//Proceedings of 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). Piscataway: IEEE Press, 2019: 16-23.

[30] WANG J, NIE X H, XIA Y, et al. Cross-view action modeling, learning, and recognition[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 2649-2656.

[31] XIA L, CHEN C C, AGGARWAL J K. View invariant human action recognition using histograms of 3D joints[C]//Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2012: 20-27.

[32] DU Y, WANG W, WANG L. Hierarchical recurrent neural network for skeleton based action recognition[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1110-1118.

[33] XIAO Y, CHEN J, WANG Y C, et al. Action recognition for depth video using multi-view dynamic images[J]. Information Sciences, 2019, 480: 287-304.

[34] YUN K, HONORIO J, CHATTOPADHYAY D, et al. Two-person interaction detection using body-pose features and multiple instance learning[C]//Proceedings of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2012: 28-35.

[35] ZHANG S Y, LIU X M, XIAO J. On geometric features for skeleton-based action recognition using multilayer LSTM networks[C]//Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2017: 148-157.

[36] ZHU W T, LAN C L, XING J L, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 3697-3703.

[作者简介]

王洪雁（1979– ），男，河南南阳人，博士，浙江理工大学特聘教授、硕士生导师，主要研究方向为阵列信号处理、机器视觉等。

袁海（1996– ），男，辽宁锦州人，大连大学硕士生，主要研究方向为图像处理、动作识别等。