中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2024)05-1421-13

论文引用格式:Deng G S, Ding W W, Yang C and Ding C Y. 2024. Gesture recognition by combining spatio-temporal mask and spatial 2D position encoding. Journal of Image and Graphics, 29(05):1421-1433(邓淦森, 丁文文, 杨超, 丁重阳. 2024. 结合时空掩码和空间二维位置编码的手势识别. 中国图象图形学报, 29(05):1421-1433)[DOI:10.11834/jig.230379]

结合时空掩码和空间二维位置编码的手势识别

邓淦森[1],丁文文[1*],杨超[1],丁重阳[2]

1. 淮北师范大学数学科学学院,淮北　235000; 2. 西安电子科技大学计算机科学与技术学院,西安　710071

摘　要:**目的**　在动态手势序列特征提取时,忽略了不同动态手势手指间的相关性,是造成手势识别率不高的重要原因。针对此问题,提出了时空位置编码和掩码的方法进行手势识别,是首次对手部关节点进行空间二维位置编码。**方法**　首先,根据手部关节序列构造时空图,利用关节点平面坐标生成空间二维编码,并与时间轴的一维编码器融合,生成关节点的时空位置编码,可以有效处理空间上的异常姿态同时避免时间上的乱序问题;然后,将时空图按照人体手部生物结构进行分块,通过空间自注意力和空间掩码,获取手指与手指之间的潜在信息。采用时间维度扩张的策略,通过时间自注意力和时间掩码,捕获长时间手指序列动态演变信息。**结果**　在DHG-14/28(dynamic hand gesture 14/28)数据集上,该算法比HPEV(hand posture evolution volume)算法平均识别率高出4.47%,比MS-ISTGCN(multi-stream improved spatio-temporal graph convolutional network)算法平均识别率高出2.71%;在SHREC'17 track数据集上,该算法比HPEV算法平均识别率高出0.47%,利用消融实验证明了本文策略的合理性。**结论**　通过大量实验评估,验证了基于分块和时空位置编码构造出来的模型很好地解决了上述问题,提高了手势识别率。

关键词:手势识别;自注意力;空间二维位置编码;时空掩码;手部分块

Gesture recognition by combining spatio-temporal mask and spatial 2D position encoding

Deng Gansen[1], Ding Wenwen[1*], Yang Chao[1], Ding Chongyang[2]

1. School of Mathematical Sciences, Huaibei Normal University, Huaibei　235000, China;

2. School of Computer Science and Technology, Xidian University, Xi'an　710071, China

Abstract: **Objective** Gesture recognition often neglects the correlation between fingers and pays excessive attention to the node features, which is crucial for the low gesture recognition rate. For example, the index finger and thumb are physically disconnected, but their interaction is important for recognizing the "pinch" action. Thus, the low recognition rate is due to the inability to encode the spatial position of the hand node properly. Dividing the joint of the hand part into blocks is proposed to address the correlation between fingers. The aforementioned problem can be addressed byencoding the two-dimensional position of the joint through its projection coordinates. The authors believe that this study is the first to encode the two-dimensional position of the node in space. **Method** The spatiotemporal graph is generated from the gesture

收稿日期:2023-06-20;**修回日期**:2023-09-19;**预印本日期**:2023-09-26

*通信作者:丁文文　dww2048@163.com

基金项目:国家自然科学基金项目(62171342);安徽省教育厅自然科学重大项目(KJ2020ZD008)

Supported by: National Natural Science Foundation of China (62171342); Major Natural Science Project of Anhui Provincial Department of Education (KJ2020ZD008)

sequence. This graph contains the physical connection of the node and its temporal information. Thus, the spatial and temporal characteristics are learned using mask operations. According to the three-dimensional space coordinates of joint nodes, the two-dimensional projection coordinates are obtained, and the two-dimensional projection coordinates are inputted into the two-dimensional space position encoder, which comprises sine and cosine functions with different frequencies. The plane where the projection coordinates are located is divided into several grid cells, and the encoder comprising sine and cosine functions is calculated in each grid cell. The encoders in all grids are combined to form sine and cosine functions with different frequencies to generate the final spatial two-dimensional position code. Embedding the encoded information into the spatial features of the nodes not only strengthens the spatial structure between them but also avoids the disorder of the nodes in the movement process. Using the graph convolutional network to aggregate and embed the spatial encoded node and neighbor features, the spatiotemporal graph features after the graph convolution are inputted into the spatial self-attention module to extract the inter-finger correlation. Taking each finger as the research object, the distribution of nodes in the spatiotemporal graph is divided into blocks according to the biological structure of the human hand. Each finger through a linear learnable change to generate the eigenvector of the finger query (Q), key (K), value (V). The self-attention mechanism is then used to calculate the correlation between fingers in each frame of the space-time graph, the correlation weight between fingers is obtained by combining the spatial mask matrix, and each finger feature is updated. While updating the finger features, the spatial mask matrix is used to disconnect the time relationship between fingers in the spatiotemporal graph, avoiding the influence of time dimension on the spatial correlation weight matrix. The time self-attention module is similarly used to learn the timing features of fingers in the spatiotemporal graph. First, temporal sequence embedding is conducted for each frame through temporal one-dimensional position coding to obtain the temporal sequence information of each frame during model learning. The time dimension expansion strategy is used to fuse the features of the two adjacent frames to capture the interframe correlation at a long distance. A learnable linear change then generates a feature vector query (Q), key (K), and value (V) for each frame. Finally, the self-attention mechanism is utilized to calculate the correlation between each frame in the space-time graph. Simultaneously, the correlation weight matrix between frames in the space-time graph is obtained by combining the time mask matrix, and the features of each frame are updated. Updating the features of each frame also uses the temporal mask matrix to avoid the influence of spatial dimension on the temporal correlation weight matrix. The fully connected network, ReLU activation function, and layer normalization are added to the end of each attention module to improve the training efficiency of the model, and the model finally outputs the learned feature vector for gesture recognition. **Result** The model is tested on two challenging datasets: DHG-14/28 and SHREC'17 track. The experimental results show that the model achieves the best recognition rate on DHG-14/28, which is 4.47% and 2.71% higher than the HPEV and the MS-ISTGCN algorithms, respectively, on average. On the SHREC'17 track dataset, the algorithm is 0.47% higher than the HPEV algorithm on average. The ablation experiment proves the need of two-dimensional location coding in space. The experimental test shows that the model has the best recognition rate when node features are 64 dimensions and the number of self-attention head is 8. **Conclusion** Numerous experimental evaluations verified that the network model constructed by the block strategy and spatial two-dimensional position coding not only improves the spatial structure of the nodes but also enhances the recognition rate of gestures using the self-attention mechanism to learn the correlation between non-physically connected fingers.

Key words: gesture recognition; self-attention; spatial two-dimensional position coding; spatio-temporal mask; hand segmentation

0 引 言

随着计算机视觉的发展,人机交互方式越来越广泛,在自动驾驶、虚拟现实等方面利用机器识别人体动作的技术也逐渐成熟。手势作为日常生活中传递简单信息的重要方式,在动作识别方面具有重要的研究意义。目前,对手势的识别方法主要分为静态手势和动态手势(Rautaray 和 Agrawal,2015;Cheng 等,2016)。静态手势利用静止的单帧图像表示,而动态手势是通过一系列手部动作序列表示,如挥手、握手等。对手部动作识别的数据模式为基于 RGB

图像模式和基于骨架序列模式。RGB图像模式从图像中提取特征进行手势识别(Molchanov等,2015;Wang等,2015);骨架序列模式通过手部关节点三维坐标表示骨架结构,不仅可以提供准确的人体结构信息,而且还提供了骨骼连接关系及人体骨骼长度等先验知识(Caputo等,2018;Oreifej和Liu,2013;Chen等,2017)。骨架序列数据模式克服了RGB图像数据模式易受外界环境干扰,如光照强弱、视觉差异等造成的识别困难。因此基于骨架序列模式的手势识别研究一直很活跃。

传统基于骨架进行识别的方法通常利用手工制作的特征描述符提取手部空间信息(Ohn-Bar和Trivedi,2013;Caputo等,2018;Oreifej和Liu,2013)。de Smedt等人(2016)基于骨架序列数据模式设计了有效的描述符,用于表示手的几何形状、手腕的旋转以及手的方向,每个描述符都被编码在傅里叶向量中,即通过使用高斯混合模型(gaussian mixture model,GMM)汇集局部图像特征而获得的表示,并采用时间金字塔对时间特征进行编码,最后通过线性支持向量机对手势进行分类。然而,这种手工特征的方法较为局限,对手势识别的效果依赖于手工制作特征的合理性,也无法更好地建模时空更深层次的信息。

随着深度学习快速发展,端到端学习方式得到许多研究者关注。通过将骨架关节点信息直接输入到网络进行手部动作识别和预测,减少了手动提取特征带来的烦琐和错误。卷积神经网络(convolutional neural network,CNN)、循环神经网络(recurrent neural network,RNN)和图卷积神经网络(graph convolution neural network,GCN)是基于深度学习方法进行手势识别研究的3个主要网络。CNN是一种前馈神经网络,利用卷积结构从数据中提取特征完成下游任务(Shiri等,2023),RNN是一类具有内部记忆的深度学习网络,能够捕获序列间依赖关系进行预测任务等(Fang等,2021)。Núñez等人(2018)提出将手关节的3个维度与图像的3个通道对应,将手骨骼序列排列成RGB图像输入CNN提取序列的空间和时间特征。Devineau等人(2018)提出了一种并行卷积网络PCNN(parallel CNN),对手骨骼关节的位置序列采用并行卷积,并通过特定通道进行特征提取,最后经过全连接网络进行动作分类。Chen等人(2017)提出使用RNN模型来学习手势的运动,利用手指运动过程中旋转和平移两个特征来表示整体运动,然后通过双向RNN进行手势识别。由于CNN无法处理拓扑结构的数据,这些方法并不能有效地学习手关节的空间特征和时空上下文信息,因此衍生出GCN方法聚合手骨架关节点的邻居信息。手骨架序列包含时间动态信息,仅对骨架的空间结构信息学习是不够的,Si等人(2019)首次将时空图卷积网络(spatio-temporal graph convolutional network,ST-GCN)(Yan等,2018)应用于动态手势识别任务,打破以往只专注于空间维度的限制,并且为了更加精细地描述关节之间的联动作用,提出通过运动学链接来建立3种类型的边。Li等人(2019)为了提取关节点之间的潜在联系,通过3种类型的边构造骨架序列的拓扑结构,并且利用关节点的相对坐标使GCN学习不依赖于关节点起始位置的动作特征。然而,构造固定的时空图不利于处理不同手势中手指间的空间关系,也无法充分建模长距离关节点的时间关系。Ding等人(2022)提出新型的时间片段图卷积网络(time segment graph convolution network,TS-GCN),首先将整个序列划分为几个子序列,然后对每个子序列应用GCN来捕捉动态信息,在时域上通过对齐运动特征来提高模型的适应能力,经过实验证明了所提网络的有效性。

Transformer模型提出的自注意力机制(Vaswani等,2017)使得模型更加关注提升长距离建模能力,实现模型自我学习和更新。随着自注意力机制应用于各种自然语言处理任务中,例如机器翻译、问答系统等,使自注意力机制成为深度学习领域重要的技术。

注意力机制也广泛地应用于其他任务中,例如,Chen等人(2019)和Si等人(2019)已经将注意力机制应用于动态手势识别任务。Chen等人(2019)提出了动态图时空自注意(dynamic graphs spatio-temporal attention)模型,利用动态手势序列构建时空图,通过自注意机制学习动态手势序列中关节点之间的空间结构和时间动态演变对时空图进行更新。Si等人(2019)基于长短期记忆网络(long short term memory,LSTM)提出了一种注意增强图卷积长短期记忆网络(attention graph convolutional LSTM,AGC-LSTM),用于选择有区别的空间信息。Shi等人(2021)提出数据解耦思想,不仅将数据解耦成4种流,而且把动作解耦到空间和时间两个维度,然

后通过位置编码为每个关节点提供唯一的标记，最后利用完全注意机制建模关节点之间的空间结构和时间信息。Song等人(2022)提出了一种新型的多流改进的时空图卷积网络(multi-stream improved spatio-temporal graph convolutional network，MS-ISTGCN)用于动态手势识别，该网络采用自适应空间图卷积学习手关节之间的关系，并提出扩展时间图卷积实现从短时间信息到长时间信息的特征提取，同时利用注意机制使模型关注关键特征。Li等人(2023)提出多层次聚合网络(multi-view hierarchical aggregation network，MVHAN)，首先用CNN提取手骨架多视图特征，然后利用层次注意体系结构和全局上下文建模融合多视图特征，将融合后的特征用于手势识别。姜叔晏等人(2022)通过嵌入在注意力机制框架中的特征融合模块获取时空依赖信息，改善了图卷积骨架行为识别方法的分类效果。何伟和潘晨(2022)结合通道注意力机制和空间注意力机制聚合深层和浅层的特征信息，便于处理不同层次特征的传递和聚合。Miah等人(2023)提出基于多分支的注意力图(multi-branch attention based graph，MBABG)模型，应用空间注意模块和时间注意模块提取基于骨骼的所有可能类型的特征，并利用深度网络得到一般特征，最后将所有特征连接输入连接层进行手势识别。

上述方法在提取手部时空特征时默认手部关节点带有周期性，利用时间位置编码对手关节嵌入身份信息，这显然是不合理的，而且上述模型只关注手部关节特征，忽略了手势动作过程中手指间的潜在信息，导致手部动作识别率较低。

结合时空掩码和空间二维位置编码的手势识别结构如图1所示，根据手骨架动态序列的时空图获取关节点投影坐标进行空间二维位置编码后，利用自适应图卷积来聚合手关节及其相邻节点特征。空间自注意模块首先根据人体手部生物结构对时空图进行分块得到基于手指$\boldsymbol{X}^l$的时空图，经过线性映射得到每个手指可学习的向量值$\boldsymbol{V}_{\mathrm{S}}$，查询$\boldsymbol{Q}_{\mathrm{S}}$，键$\boldsymbol{K}_{\mathrm{S}}$，然后，由自注意力机制得到手指间空间相关性权重$\boldsymbol{U}_{\mathrm{S}}$，结合空间掩码矩阵$\boldsymbol{M}_{\mathrm{S}}$同时掩盖时间维度$(1-\boldsymbol{M}_{\mathrm{S}})\times\eta$计算时空图中手指间的空间相关性$\boldsymbol{W}_{\mathrm{S}}$，最后更新手指空间特征$\boldsymbol{W}_{\mathrm{S}}\boldsymbol{V}_{\mathrm{S}}$。时间自注意模块首先对时空图进行维度扩张并嵌入时间一维位置编码，同样经过线性变换得到序列每帧的特征向量值$\boldsymbol{V}_{\mathrm{T}}$，查询$\boldsymbol{Q}_{\mathrm{T}}$，键$\boldsymbol{K}_{\mathrm{T}}$，然后与空间自注意模块相同通过自注意力机制和时间掩码矩阵$\boldsymbol{M}_{\mathrm{T}}$，最后更新每帧特征$\boldsymbol{W}_{\mathrm{T}}\boldsymbol{V}_{\mathrm{T}}$。在空间和时间自注意力模块中加入全连接层、ReLU(rectified linear unit)激活函数层和归一化层，目的是提高模型收敛速度和防止模型产生梯度爆炸，经过模型对手势时空图特征提取输出与动作类别数相同的向量。

本文的主要贡献如下：1)为了更加充分地提取不同手势的时空特征，空间维度中将手势动作序列的时空图按照人体手部生物结构进行分块表示；时间维度中对时空图进行相邻帧扩张，提高模型对时间特征的提取范围。2)首次采用空间二维位置编码和时间编码操作对手部三维关节点进行时空位置编码。3)对模型添加时空掩码掩盖不同维度的边连接，使模型专注学习时空特征。

1 基础知识

1.1 图卷积网络

对于存储形式为图的数据，如分子网络、社交网络以及引用论文网络等，无法直接利用RNN和CNN等网络进行建模，因此图卷积网络建模拓扑结构的数据引起了许多人的关注。给定图结构$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中，$\boldsymbol{V}=\{V_i\,|\,i=1,\cdots,N\}$表示图中所有节点，$\boldsymbol{E}=\{(V_{is},V_{ie})\,|\,V_{is},V_{ie}\in\boldsymbol{V}\}$表示图中边连接，如式(1)所示，通过图卷积网络建模图中节点的邻居信息，具体为

$$\boldsymbol{H}^{(l+1)}=\sigma(\boldsymbol{A}(\theta)\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{1}$$

式中，$\boldsymbol{A}(\theta)\in\boldsymbol{R}^{N\times N}$为参数化的邻接矩阵，通过模型自动学习图中任意两个点之间的连接关系，N为图中节点的个数，$\boldsymbol{H}^{(l)}$代表图中第l层的节点特征，$\boldsymbol{W}^{(l)}$为第l层特征对应的权重矩阵，$\sigma(\cdot)$为激活函数。

1.2 时空位置编码

Vaswani等人(2017)在Transformer模型中提出适用于周期性数据的位置编码，其关键思想是文字建模。这种编码方式有助于在手指序列中对每个手指的进行时间位置编码，即

$$PE(pos,i)=\begin{cases}\sin\dfrac{pos}{10\,000^{\frac{2i}{d}}}\\[2ex]\cos\dfrac{pos}{10\,000^{\frac{2i}{d}}}\end{cases} \tag{2}$$

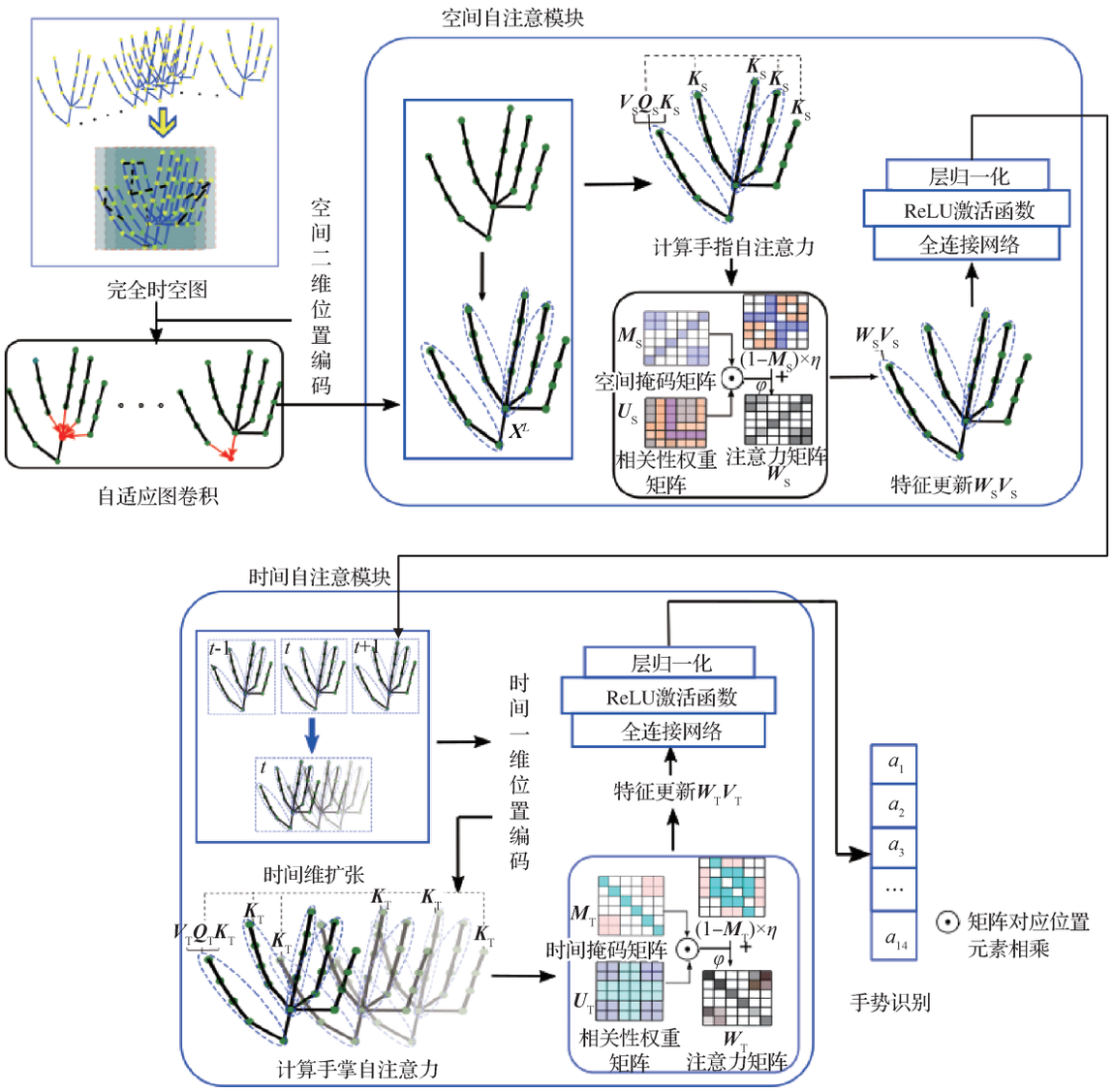


图 1　网络模型结构图

Fig. 1　Network model structure diagram

式中，$0 < i < d$，i 为编码特征维度，pos 为手指所在位置，分别用于序列中偶数和奇数位置上的编码。正弦和余弦函数具有周期性，因此对于具有周期性的手指序列中 $pos + k$ 的位置可以表示成 pos 位置的线性变化，k 表示在序列中的偏移量，方便模型学习手指与手指之间的相对位置关系。

图卷积网络（GCN）为拓扑结构的数据提供了一种强大且可扩展的解决方案。然而，对于手关节点在缺乏关于数据几何结构上下文的情况下，通常依赖手关节点之间的欧氏距离来构建图结构，这样容易造成关节点在空间位置和上下文信息的丢失。Mai 等人（2020）提出对地理坐标的空间位置编码方法，受此启发，在得到手部关节点的平面坐标 $\boldsymbol{X}$ 后，对手部关节点定义二维空间位置编码，具体为

$$P = NN(E(\boldsymbol{X})) \tag{3}$$

式中，$E(\boldsymbol{X}) = [E_0(\boldsymbol{X}); \cdots; E_s(\boldsymbol{X}); \cdots; E_{S-1}(\boldsymbol{X})]$ 由不同频率的正余弦函数构成，S 为空间划分网格单元总数，$s = 0, 1, \cdots, S-1$，$NN(\cdot)$ 是包含 ReLU 的全连接层。定义两两夹角为 $2\pi/3$ 的二维单位向量，$\boldsymbol{\alpha}_1 = [1, 0]^{\mathrm{T}}$，$\boldsymbol{\alpha}_2 = [-1/2, \sqrt{3}/2]^{\mathrm{T}}$，$\boldsymbol{\alpha}_3 = [-1/2, -\sqrt{3}/2]^{\mathrm{T}}$，在每个网格单元中，$E_s(\boldsymbol{X})$ 的计算式为

$$E_s(\boldsymbol{X}) = \left[\cos\left(\frac{\langle \boldsymbol{X}, \boldsymbol{\alpha}_j \rangle}{\sigma_{\min} g^{\frac{s}{S-1}}}\right); \sin\left(\frac{\langle \boldsymbol{X}, \boldsymbol{\alpha}_j \rangle}{\sigma_{\min} g^{\frac{s}{S-1}}}\right)\right] \tag{4}$$

式中，$\langle \cdot \rangle$ 表示矩阵相乘，$j = 1, 2, 3$，$g = \sigma_{\max}/\sigma_{\min}$，$\sigma_{\max}$，$\sigma_{\min}$ 为网格单元的最大、最小值。

1.3　自注意力机制

自注意力机制（Vaswani 等，2017）更擅长捕捉数

据自身内部的相关性，降低对其他外部信息的需求和依赖，通过点乘法计算任意两个数据之间的相关性实现模型的自我更新。注意力机制的输入包括维度为d_k的查询$\boldsymbol{Q}$、键$\boldsymbol{K}$和值$\boldsymbol{V}$，通过计算查询与键的点乘并应用归一化指数函数得到数据之间的相关程度，具体为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

最常用的计算注意力方式是加法计算和点乘法（乘法）计算，加法计算使用带有前馈网络的兼容性函数来计算单一隐藏层，虽然两者在理论上的复杂性相似，但是点乘法计算在实践中速度更快、空间效率更高，因为它可以用高度优化的矩阵乘法代码。当注意力机制输入维度d_k值比较小时，这两种机制计算注意力的表现相似；但是对于较大的d_k时，加法计算函数的表现优于乘法计算函数。d_k的值越大使得乘积也会越大，造成将归一化指数函数推向梯度极小的区域，因此用$\sqrt{d_k}$来抵消该影响。

1.4 时空掩码操作

掩码操作利用掩码矩阵遮盖不需要学习的特征维度，不仅使模型更加专注于当前维度的学习，而且提高了矩阵的计算效率（Chen等，2019），具体为

$$\boldsymbol{W}=\varphi(\boldsymbol{U}\odot\boldsymbol{M}+(1-\boldsymbol{M})\times\eta) \tag{6}$$

式中，$\boldsymbol{U}=\dfrac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}$，$\odot$为对应元素的乘积，$\boldsymbol{M}$为掩码矩阵，将需要学习的特征维度对应位置的元素保留为1，其余位置元素设为0，φ为归一化指数函数，η为趋近于负无穷的值，$\boldsymbol{W}$为加入掩码后的注意力矩阵，时空掩码过程如图2所示，通过空间掩码矩阵保留时空图中关节点的空间结构，断开关节点时序上的连接，避免时间动态相关对空间相关性造成影响；同样，应用时间掩码矩阵避免空间相关性对时间动态相关性的影响。

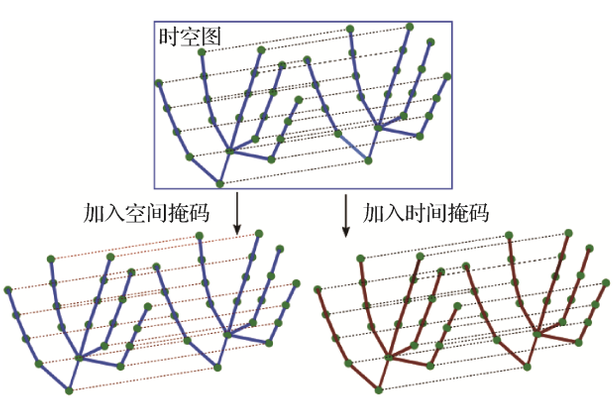


图2 时空掩码过程

Fig. 2 Spatio-temporal masking process

2 基于二维编码的分块自注意网络

给定手骨架动态序列$\boldsymbol{L}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_T]$，$T$为骨架序列的帧数，$\boldsymbol{X}=[x_1,x_2,\cdots,x_N]\in\mathbf{R}^{N\times 3}$为每帧手骨架空间特征，$N$为三维关节点个数，每一帧关节点位置如图3所示。根据Chen等人（2019）的方法构造统一完全时空图$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中，$\boldsymbol{V}=\{v_{(i,t)}\,|\,i=1,2,\cdots,N,t=1,\cdots,T\}$表示所有关节点集，$\boldsymbol{E}=\{(v_{(t,i)},v_{(t,j)}),(v_{(t,i)},v_{(k,j)})\}$表示时空图的边集，$(v_{(t,i)},v_{(t,j)})$表示每个关节点与其所在空间的其他关节点的连接，$(v_{(t,i)},v_{(k,j)})$表示将前一帧的每个关节点与后一帧所有关节点的连接。

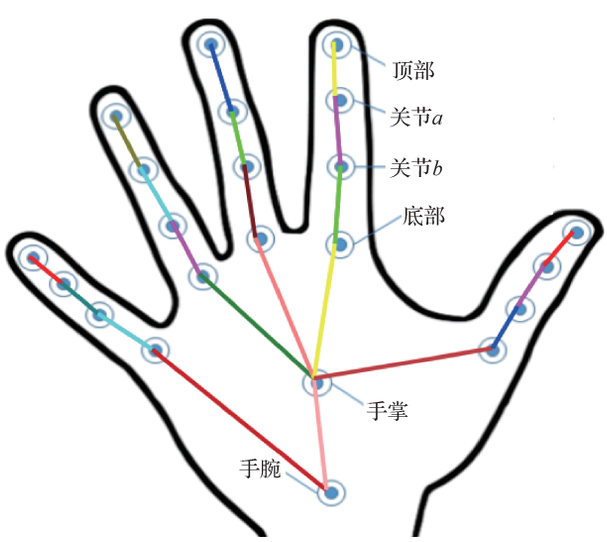


图3 手势关节位置图

Fig. 3 Gesture joint position

2.1 空间位置编码

给定手部关节点$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]\in\mathbf{R}^{N\times 3}$，$N$为关节点个数，$\boldsymbol{x}_i=(x_1,x_2,x_3)$为关节点三维坐标，将三维关节点投影到$xy$平面得到关节点平面坐标$\boldsymbol{X}_{xy}=[\boldsymbol{x}_i(x_1,x_2)]\in\mathbf{R}^{N\times 2}$，对平面坐标$\boldsymbol{X}_{xy}$进行空间二维位置编码，具体为

$$\boldsymbol{P}=NN(E(\boldsymbol{X}_{xy})) \tag{7}$$

式中，$E(\boldsymbol{X}_{xy})=[E_0(\boldsymbol{X}_{xy});\cdots;E_s(\boldsymbol{X}_{xy});\cdots;E_{S-1}(\boldsymbol{X}_{xy})]$，

$$E_s(\boldsymbol{X}_{xy})=\left[\cos\left(\frac{\langle \boldsymbol{X}_{xy},\boldsymbol{\alpha}_j\rangle}{\sigma_{\min}g^{\frac{s}{S-1}}}\right);\ \sin\left(\frac{\langle \boldsymbol{X}_{xy},\boldsymbol{\alpha}_j\rangle}{\sigma_{\min}g^{\frac{s}{S-1}}}\right)\right],\quad s=0,1,\cdots,S-1,j=1,2,3$$

对 $\boldsymbol{X}$ 嵌入位置编码，具体为

$$\boldsymbol{X}=\boldsymbol{X}+\boldsymbol{P} \tag{8}$$

将 $\boldsymbol{X}$ 作为关节点的初始特征输入GCN，如式(1)所示，聚合关节点及其邻居节点特征，提升每个关节点的特征维度，具体为

$$\boldsymbol{X}^L=\sigma(\boldsymbol{A}(\theta)\boldsymbol{X}\boldsymbol{W}) \tag{9}$$

式中，更新后的每帧关节点特征为 $\boldsymbol{X}^L\in\mathbf{R}^{N\times d}$，$\boldsymbol{W}^L\in\mathbf{R}^{3\times d}$，$d>3$，$d$ 为关节点更新后的特征维数。

2.2 对手部分块

在动态手势识别过程中，仅考虑手部关节点的物理连接，而不考虑各手指之间的关联性，是造成手部动作识别率低的重要原因。针对此问题，本文采用对手骨架分块策略，将骨架按照人体手部生物结构分块，其中，掌心处的关节被分配到4个手指内，为了增强主要动作手指间的联系，最后将剩下的腕关节与拇指关节视为一个手指，如图4所示。同理，将时空图分为5块，复制掌心节点特征到4个手指特征分组中，分块后每帧手骨架特征为 $\boldsymbol{Y}^{\text{part}}=[\boldsymbol{X}^L_{\text{part},1},\cdots,\boldsymbol{X}^L_{\text{part},5}]\in\mathbf{R}^{5\times(\lfloor\frac{N}{5}\rfloor\times d)}$，$\boldsymbol{X}^L_{\text{part},i}\in\mathbf{R}^{\lfloor\frac{N}{5}\rfloor\times d}$ 为每个手指特征。

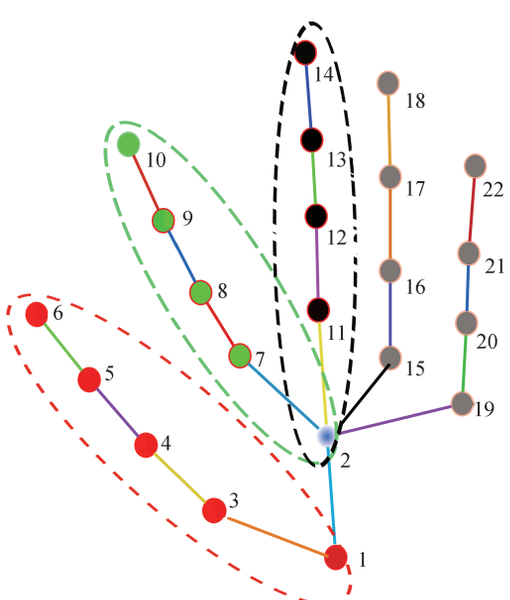


图 4 手骨架分块

Fig. 4 Hand skeleton in parts

2.3 空间自注意力模块

为了学习每个手指之间的潜在信息，对手指特征应用可训练的线性变化得到查询矩阵 $\boldsymbol{Q}_{\text{S}}$、键矩阵 $\boldsymbol{K}_{\text{S}}$ 和值矩阵 $\boldsymbol{V}_{\text{S}}$，然后利用注意力机制来学习手指间相关性，具体为

$$\begin{cases}\boldsymbol{Q}_{\text{S}}=\boldsymbol{Y}^{\text{part}}\boldsymbol{W}_{Q_{\text{S}}}\\ \boldsymbol{K}_{\text{S}}=\boldsymbol{Y}^{\text{part}}\boldsymbol{W}_{K_{\text{S}}}\\ \boldsymbol{V}_{\text{S}}=\boldsymbol{Y}^{\text{part}}\boldsymbol{W}_{V_{\text{S}}}\end{cases} \tag{10}$$

式中，权重矩阵 $\boldsymbol{W}_{Q_{\text{S}}},\boldsymbol{W}_{K_{\text{S}}},\boldsymbol{W}_{V_{\text{S}}}\in\mathbf{R}^{(h\times d_w)\times(\lfloor\frac{N}{5}\rfloor\times d)}$ 在所有手指之间共享，d_w 为每个自注意力头的维度，h 为自注意力头的个数。通过点乘法计算任意两个手指间的相关性矩阵 $\boldsymbol{U}_{\text{S}}\in\mathbf{R}^{5\times5}$，具体为

$$\boldsymbol{U}_{\text{S}}=\frac{\boldsymbol{Q}_{\text{S}}\boldsymbol{K}_{\text{S}}^{\text{T}}}{\sqrt{5\times h\times d_w}} \tag{11}$$

对 $\boldsymbol{U}_{\text{S}}$ 应用掩码矩阵防止空间与时间特征发生信息混乱，同时达到监督模型学习的目的，具体为

$$\boldsymbol{W}_{\text{S}}=\varphi(\boldsymbol{U}_{\text{S}}\odot\boldsymbol{M}_{\text{S}}+(1-\boldsymbol{M}_{\text{S}})\times\eta) \tag{12}$$

式中，$\boldsymbol{M}_{\text{S}}\in\mathbf{R}^{5\times5}$ 为空间掩码矩阵，将表示空间关系的值保留为1，时间关系的值设为0，同时将表示时间的边对应权重值设为 η，经过指数归一化激活函数得到加入掩码后的相关性权重 $\boldsymbol{W}_{\text{S}}$，并与手指对应的值矩阵 $\boldsymbol{V}_{\text{S}}$ 进行加权，经过全连接层 $NN(\cdot)$ 将特征映射到初始维度，利用 $LN(\cdot)$ 对网络的输出进行归一化提高模型学习效率，最终得到空间注意模块的输出，具体为

$$\boldsymbol{Y}=LN(ReLU(NN(\boldsymbol{W}_{\text{S}}\boldsymbol{V}_{\text{S}}))) \tag{13}$$

2.4 时间扩张策略

仅考虑每一帧内手指与其相邻两帧之间的关系并不能捕捉到手指远距离的潜在相关性，这是造成动态手势识别不精确的重要原因。针对此问题，本文提出在时间维度利用扩张策略，进一步提高时间注意模块对手指长远距离动态信息的提取能力。将空间注意模块输出的时空图进行时间扩张，即把相邻3帧 $t-1,t,t+1$ 以及 $t,t+1,t,t+2$ 进行合并，具体操作如图5所示，合并前时空图的每一帧特征为 $\boldsymbol{Y}\in\mathbf{R}^{5\times(\lfloor\frac{N}{5}\rfloor\times d)}$，合并后的特征为 $\tilde{\boldsymbol{Z}}_j\in\mathbf{R}^{(3\times5)\times(\lfloor\frac{N}{5}\rfloor\times d)}$，具体为

$$\tilde{\boldsymbol{Z}}_j=[\boldsymbol{Y}_j,\boldsymbol{Y}_{j+1},\boldsymbol{Y}_{j+2}]\in\mathbf{R}^{(3\times5)\times(\lfloor\frac{N}{5}\rfloor\times5)} \tag{14}$$

式中，$\tilde{\boldsymbol{Z}}=\{\tilde{\boldsymbol{Z}}_j\}\in\mathbf{R}^{\tau\times15\times(\lfloor\frac{N}{5}\rfloor\times d)}$，$\tau=T-2$。

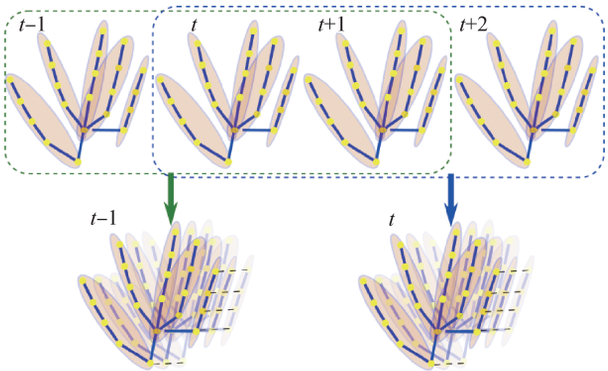


图 5　时间维度扩张

Fig. 5　Time dimension extension

2.5　时间位置编码

将合并后的时空图 $\tilde{\boldsymbol{Z}} \in \mathbf{R}^{\tau \times 15 \times (\lfloor \frac{N}{5} \rfloor \times d)}$ 作为输入进行位置编码，对每帧中所有手指从1到15进行编号，生成[1；2；…；15；16；…；15×T]列向量，列向量中每个元素对应每个手指的位置pos，进行位置嵌入，具体为

$$\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{Z}} + PE(pos, i) \tag{15}$$

式中，$PE(pos, i) \in \mathbf{R}^{\tau \times 15 \times (\lfloor \frac{N}{5} \rfloor \times d)}$，$pos$表示手指所在位置，$m = (\lfloor \frac{N}{5} \rfloor \times d), 1 < i < m$。

2.6　时间注意模型

时间注意模型计算方法与空间注意模型对称，对时空图特征 $\tilde{\boldsymbol{Z}} \in \mathbf{R}^{\tau \times 15 \times (\lfloor \frac{N}{5} \rfloor \times d)}$ 同样进行可训练的线性变化，得到查询矩阵$\boldsymbol{Q}_{\mathrm{T}}$、键矩阵$\boldsymbol{K}_{\mathrm{T}}$以及值矩阵$\boldsymbol{V}_{\mathrm{T}}$，具体为

$$\begin{cases} \boldsymbol{Q}_{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{part}} \boldsymbol{W}_{Q_{\mathrm{T}}} \\ \boldsymbol{K}_{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{part}} \boldsymbol{W}_{K_{\mathrm{T}}} \\ \boldsymbol{V}_{\mathrm{T}} = \boldsymbol{Y}^{\mathrm{part}} \boldsymbol{W}_{V_{\mathrm{T}}} \end{cases} \tag{16}$$

式中，权重矩阵 $\boldsymbol{W}_{Q_{\mathrm{T}}}, \boldsymbol{W}_{K_{\mathrm{T}}}, \boldsymbol{W}_{V_{\mathrm{T}}} \in \mathbf{R}^{(h \times d_w) \times (\lfloor \frac{N}{5} \rfloor \times d)}$ 在所有帧之间共享。利用点乘法计算每个手指在时间维度的相关性矩阵$\boldsymbol{U}_{\mathrm{T}}$，具体为

$$\boldsymbol{U}_{\mathrm{T}} = \frac{\boldsymbol{Q}_{\mathrm{T}} \boldsymbol{K}_{\mathrm{T}}^{\mathrm{T}}}{\sqrt{\tau \times 15 \times h \times d_w}} \tag{17}$$

将时间掩码矩阵$\boldsymbol{M}_{\mathrm{T}}$作用于$\boldsymbol{U}_{\mathrm{T}}$，使时间注意模块专注于学习时间维度，利用归一化指数函数φ，得到时间注意力矩阵$\boldsymbol{U}_{\mathrm{T}}$，具体为

$$\boldsymbol{W}_{\mathrm{T}} = \varphi(\boldsymbol{U}_{\mathrm{T}} \odot \boldsymbol{M}_{\mathrm{T}} + (1 - \boldsymbol{M}_{\mathrm{T}}) \times \eta) \tag{18}$$

$\boldsymbol{M}_{\mathrm{T}} \in \mathbf{R}^{(\tau \times 15) \times (\tau \times 15)}$ 实现了将表示时间连接的边值保留为1、空间连接的边值为0，并在计算时间注意力矩阵时把时空图中表示空间的边权重值设为η（一个负无穷的值）。利用学习到的手指时间相关性与对应的每帧特征加权，同样经过全连接层将特征映射到初始维度，利用$LN(\cdot)$对网络的输出进行归一化，最终得到时间注意模型的输出 $\boldsymbol{Z} \in \mathbf{R}^{\tau \times 15 \times (\lfloor \frac{N}{5} \rfloor \times d)}$，具体为

$$\boldsymbol{Z} = LN(ReLU(NN(\boldsymbol{W}_{\mathrm{T}} \boldsymbol{V}_{\mathrm{T}}))) \tag{19}$$

2.7　模型融合

网络如何与模型进行有效对接是一个不可忽略的问题。将数据集$\boldsymbol{D}$的负对数似然损失函数用于测量真实标签$\boldsymbol{C}$和预测结果$\boldsymbol{C}^{(i)}$之间的差异，具体为

$$L(\boldsymbol{W}_{\mathrm{S}}, \boldsymbol{W}_{\mathrm{T}}, \boldsymbol{D}) = \frac{1}{|\boldsymbol{D}|} \sum_{i=0}^{|\boldsymbol{D}|} \log(\varphi(\boldsymbol{C}^{(i)} | \boldsymbol{X}^{(i)}, \boldsymbol{W}_{\mathrm{S}}, \boldsymbol{W}_{\mathrm{T}})) \tag{20}$$

式中，$|\boldsymbol{D}|$为手部动作数据集样本大小，$\boldsymbol{X}^{(i)}$为第i个动作序列特征，φ为归一化指数函数，整个模型与嵌入式注意机制不同。因此，可以使用反向传播来最小化损失函数，整个框架可以端到端进行训练。

3　实验测试

为了证明模型的有效性，在DHG-14/28（dynamic hand gesture 14/28）数据集和SHREC’17 Track数据集上进行测试，实验测试在显卡为NVIDIA GeForce RTX 3080上进行，当模型在执行50次迭代没有提升表现时就停止训练。

3.1　DHG-14/28数据集

DHG-14/28数据集（De Smedt等，2016）包含使用一根手指和整只手两种方式执行的14种手势序列，每个手势由20位实验者参与，每位实验者利用右手对上述两种方式的14类动作执行5次，一共形成2 800个动作序列。每一帧序列包含一幅深度图像、22个手关节，在三维世界空间的坐标构成一个手骨架，每个动作的长度从20～50帧。

为了使实验公平，针对每个动作序列执行均匀采样，提取8帧作为动作序列，对提取的每一帧数据进行平移、添加噪声、缩放、时间插值（Núñez等，2018），并利用第1帧的手掌位置来减去每个骨架序列以进行对齐（de Smedt等，2017）。

设置训练数据和测试数据如下：使用交叉验证的方法进行实验，将14个类别的手势全部输入到训

练网络模型，取其中一名实验者的手势序列作为测试集，将其余19名实验者的手势序列作为训练集。将关节点特征经过图卷积网路提升到64维，空间自注意力头数h为8，注意力空间维数d_w为32，本文模型与其他方法的性能对比表现如表1所示。本文模型在14个手势类别中获得了98.17%的识别率，在28个手势类别中获得了92.20%的识别率，它的性能超越了以往的方法。注意：14个手势类别比28个手势类别的识别效果更加显著，本文认为空间二维位置编码有效地编码了手关节的空间结构，因此，模型可以更好地区分不同类别的动作。

表1　在DHG数据集下所提方法与其他先进方法识别率对比结果

Table 1　Comparison results of recognition rate between our method and other advanced methods on DHG dataset

/%

方法	手势类别数	
	14	28
ST-GCN（Yan等，2018）	91.20	87.10
PCNN（Devineau等，2018）	91.28	91.28
DG-STA（Chen等，2019）	91.90	88.00
MBABG（Miah等，2023）	92.00	88.78
HPEV（Liu等，2020）	92.54	88.86
MVHAN（Li等，2023）	92.36	89.56
SP-Stream（Li等，2021）	93.10	89.82
MS-ISTGCN（Song等，2022）	93.70	91.20
本文	**98.17**	**92.20**

注：加粗字体表示各列最优结果。HPEV：hand posture evolution volume；SP-Stream：spatial perception stream。

为了研究关节点特征维数的影响，本文进行了消融实验，重新将关节点特征从3维提升到128维和32维，如表2所示，本文模型仍然具有较高的识别效果，但没有64维特征表现出色。本文认为关节点维度较高，可能造成关节点空间特征冗余使模型过拟合，维度较低模型无法充分学习空间特征。

表2　特征维度提升后的实验结果识别率

Table 2　Recognition rate ofter feature dimension on enhancennet

/%

特征映射维数	手势类别数	
	14	28
32	94.28	91.66
64	98.17	92.20
128	95.00	89.28

3.2　SHREC’17 track 数据集

SHREC’17 track 数据集（de Smedt等，2017）是由深度摄像机捕获，包含二维深度图像和22个关节的三维坐标数据，是一个极具挑战性的动态手势数据集，同样包含了由28个实验者执行的14类动作，分为14和28个手势协议。该数据集包含2 800个序列，其中有已经分好的1 960个序列用于模型训练，840个序列用于测试，每个样本手势的长度在20帧到50帧左右。同样，对SHREC’17 track 数据集中每个样本序列执行均匀采样，提取其中8 帧作为动作序列长度，经过全连接层将关节点特征维度映射为64，空间子注意力头数h为8，每个注意力空间维数d_w为32，对提取的每一帧数据同样进行平移、添加噪声、缩放、时间插值（Núñez等，2018），用之后的第1帧的手掌位置来减去每个骨架序列以进行对齐（de Smedt等，2017），并利用学习率为0.001的Adam优化器对模型进行训练。

本文模型在SHREC’17 track数据集上的表现与其他方法对比如表3所示。本文模型对14个手势类别的识别率为95.79%，在28个手势类别的识别率为92.35%。实验结果显示，本文模型与其他先进方法

表3　在SHREC’17 track数据集下所提方法与其他先进方法识别率对比结果

Table 3　Comparison results of recognition rate between our method and other advanced methods on SHREC’17 track dataset

/%

方法	手势类别数	
	14	28
STA-Res-TCN（Hou等，2018）	91.10	91.10
ST-GCN（Yan等，2018）	92.70	92.70
DG-STA（Chen等，2019）	94.40	90.70
MVHAN（Li等，2023）	94.84	92.56
HPEV（Liu等，2020）	94.90	92.30
MS-ISTGCN（Song等，2022）	96.70	94.90
MBABG（Miah等，2023）	**97.01**	**92.78**
本文	95.79	92.35

注：加粗字体表示各列最优结果。STA-Res-TCN：spatial-temporal attention residual temporal convolutional network。

相比提升效果并不显著，与DHG-14/28数据集上的识别性能相比较，对SHREC'17 track数据集进行相同的采样预处理可能进一步提高了手骨架时空结构的复杂度，不利于模型学习动态手势的时间演变。

为了研究模型在关节特征维度变化时的识别效果，将关节点特征提升为128维、32维，模型识别效果如表4所示，在关节点特征维度较高和较低的情况下，对模型的性能都有一定的影响，输入特征的维度过低时，可能手部骨架包含信息量不足以充分表示手部动作，使得模型识别率不高，若输入特征的维度较高，可能造成主要信息的稀释。

表4　特征维数映射实验结果

Table 4　Experimental results of feature dimension mapping

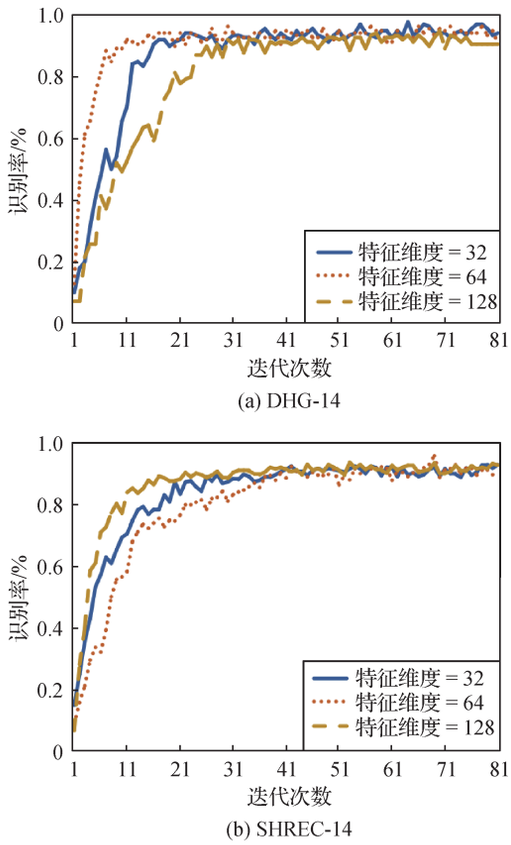
/%

特征映射维数	手势类别数	
	14	28
32	91.66	88.57
64	95.79	92.35
128	92.03	86.30

为了研究在不同特征维度下，随着训练次数增加模型的性能变化，本文考虑对模型进行80次迭代，记录模型对两个数据集的识别率，如图6所示。在DHG-14数据集关节点特征维度为64维时，模型对手势识别具有较好的性能，在同样训练次数中，模型在特征维度为128维和32维时，需要更多次迭代达到平稳识别率，这可能由于特征的稀疏或冗余造成影响。SHREC-14数据集（指SHREC'17 track中包含14个手势一类）中模型达到最好的识别性能需要更久地训练，原因可能是动作的复杂度较高，需要学习的特征更加丰富。

为了更好地说明本文模型的性能，14个动作类别的混淆矩阵如图7所示。可以看到在DHG-14数据集中，由于抓和捏两种手势高度相似，因此，在这两类动作之间存在混淆，而对于其他动作类别，模型能够准确识别。在SHREC-14数据集中，同样是对抓和捏这两类手势存在严重混淆，而且对于其他动作存在误差性分类错误，原因是模型无法清晰地学习各类动作的关键性特征。

本文研究了自注意力头数对模型性能的影响，



(a) DHG-14

(b) SHREC-14

图6　模型对DHG-14和SHREC-14不同维度的识别率

Fig. 6　Recognition rate of the model for different dimensions of DHG-14 and SHREC-14

((a) DHG-14; (b) SHREC-14)

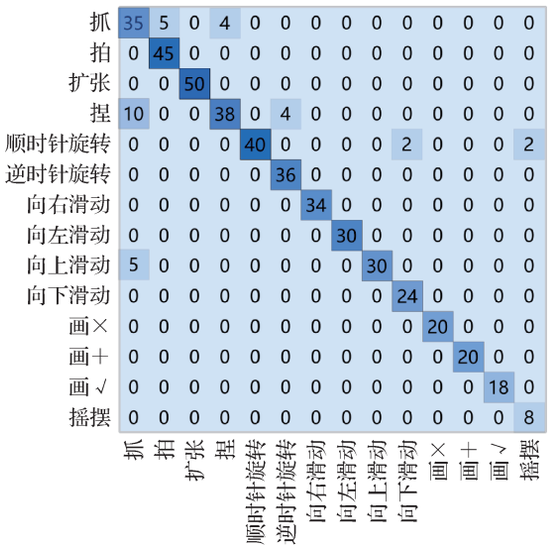
设置注意力头数为h，模型对14个动作类别的识别率，如表5所示。

最终，验证本文模型在注意力头数h为8关节点、特征维度为64时，模型的识别率达到最高，而且我们发现特征维度不发生变化时，更多的注意力头数对模型的性能没有影响。

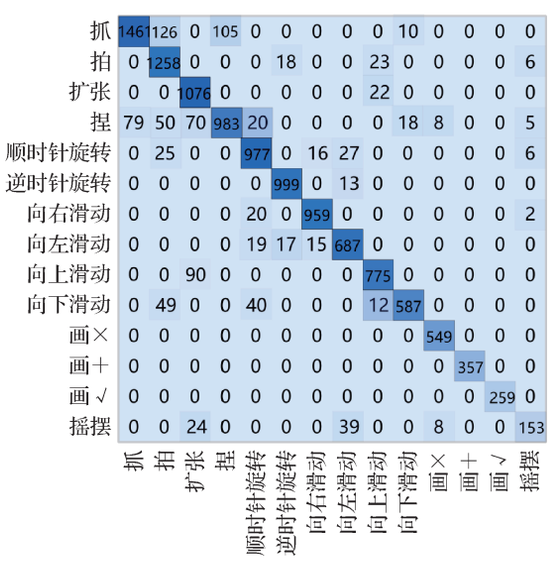
本文还验证了加入空间二维位置编码对模型性能的提升是必要的，映射向量为64维，结果如表6所示。可以看到，加入空间二维位置编码后，模型的性能明显提高，原因是二维空间位置编码使得模型获得了更多关节点的位置信息，可以提高对关节点运动特征的学习，同时验证了本文方法对提高动态手势识别效果是必要的。

4　结　论

本文首次提出了对手部关节点进行空间二维位置编码，不仅使手部关节点之间具有更强的空间结



(a) DHG-14



(b) SHREC-14

图 7　模型在 DHG-14 和 SHREC-14 数据集的混淆矩阵

Fig. 7　Confusion matrix of the model on DHG-14 dataset and SHREC-14 dataset ((a) DHG-14; (b) SHREC-14)

表 5　注意力头数对识别率的实验结果

Table 5　Experimental results of the number of attention heads versus recognition rate

/%

注意力头数	DHG-14	SHREC'17 track
6	96.42	93.27
8	98.17	95.79
10	97.85	94.79
12	97.85	94.71

构，而且防止在对关节点进行图卷积时在空间位置和上下文信息的丢失。为了弥补在对手势进行识别过程中，因忽略手指间的间接相关性造成手势识别效率较低，通过对手部关节点按照人体手部生物结构进行分块，并利用自注意机制关注非物理连接之间的优势，结合空间掩码矩阵更新手指特征。在时间维度中，针对无法得到远距离帧之间的信息的问题，对时空图采取时间维扩张，扩张后的每一帧包含更多有用信息，同样利用自注意机制和时间掩码矩阵更新每帧特征。

表 6　空间二维编码对本模型的重要性

Table 6　The importance of spatial two-dimensional coding for our model

数据集	识别率/%	
	未加入空间二维位置编码	加入空间二维位置编码
DHG_14	92.14	98.17
DHG_28	85.71	92.20
SHREC_14	91.30	95.79
SHREC_28	82.97	92.35

在模型识别性能方面，本文方法与目前最优的模型对比，在数据集 DHG-14/28 上获得了最高识别率，并通过消融实验证明了对手部关节点嵌入空间二维位置编码的必要性，以及模型经过较低训练即可达到最好性能。值得注意的是，本文在 SHREC'17 track 数据集上未能取得令人满意的结果，但是识别率与其他先进方法相比较仍然具有竞争性。

在 SHREC'17 track 数据集上识别率低的原因可能是该数据集预先经过数据加工，网络提取每一帧的特征后存在信息冗余，后续研究模型在针对不同数据集特征学习过程中提高网络的丢弃率；其次，还要验证时间维度的扩张程度对模型识别率的影响。后续研究应该考虑如何保留模型提取的原始时空特征，即设计空间自注意模块和时间自注意模块，并行对空间和时间特征进行提取。

参考文献(References)

Caputo F M, Prebianca P, Carcangiu A, Spano L D and Giachetti A. 2018. Comparing 3D trajectories for simple mid-air gesture recognition. Computers and Graphics, 73: 17-25 [DOI: 10.1016/J.CAG.2018.02.009]

Chen X H, Guo H K, Wang G J and Zhang L. 2017. Motion feature aug-

mented recurrent neural network for skeleton-based dynamic hand gesture recognition//Proceedings of 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE: 2881-2885 [DOI: 10.1109/ICIP.2017.8296809]

Chen Y X, Zhao L, Peng X, Yuan J B and Metaxas D N. 2019. Construct dynamic graphs for hand gesture recognition via spatial-temporal attention [EB/OL]. [2023-06-05]. https://arxiv.org/pdf/1907.08871.pdf

Cheng H, Yang L and Liu Z C. 2016. Survey on 3D hand gesture recognition. IEEE Transactions on Circuits and Systems for Video Technology, 26(9): 1659-1673 [DOI: 10.1109/TCSVT.2015.2469551]

de Smedt Q, Wannous H and Vandeborre J P. 2016. Skeleton-based dynamic hand gesture recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA: IEEE: 1206-1214 [DOI: 10.1109/CVPRW.2016.153]

de Smedt Q, Wannous H, Vandeborre J P, Guerry J, Le Saux B and Filliat D. 2017. 3D hand gesture recognition using a depth and skeletal dataset: SHREC'17 track//Proceedings of the Workshop on 3D Object Retrieval. Lyon, France: Eurographics Association: 33-38 [DOI: 10.2312/3dor.20171049]

Devineau G, Moutarde F, Xi W and Yang J. 2018. Deep learning for hand gesture recognition on skeletal data//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018). Xi'an, China: IEEE: 106-113 [DOI: 10.1109/FG.2018.00025]

Ding C Y, Wen S, Ding W W, Liu K and Belyaev E. 2022. Temporal segment graph convolutional networks for skeleton-based action recognition. Engineering Applications of Artificial Intelligence, 110: #104675 [DOI: 10.1016/j.engappai.2022.104675]

Fang W, Chen Y P and Xue Q Y. 2021. Survey on research of RNN-based spatio-temporal sequence prediction algorithms. Journal on Big Data, 3(3): 97-110 [DOI: 10.32604/JBD.2021.016993]

He W and Pan C. 2022. The salient object detection based on attention-guided network. Journal of Image and Graphics, 27(4): 1176-1190 (何伟, 潘晨. 2022. 注意力引导网络的显著性目标检测. 中国图象图形学报, 27(4): 1176-1190) [DOI: 10.11834/jig.200658]

Hou J X, Wang G J, Chen X H, Xue J H, Zhu R and Yang H Z. 2018. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition//Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer: 273-286 [DOI: 10.1007/978-3-030-11024-6_18]

Jiang Q Y, Wu X J and Xu T Y. 2022. M2FA: multi-dimensional feature fusion attention mechanism for skeleton-based action recognition. Journal of Image and Graphics, 27(8): 2391-2403 (姜权晏, 吴小俊, 徐天阳. 2022. 用于骨架行为识别的多维特征嵌合注意力机制. 中国图象图形学报, 27(8): 2391-2403) [DOI: 10.11834/JIG.210091]

Li S C, Liu Z Y, Duan G F and Tan J R. 2023. MVHANet: multi-view hierarchical aggregation network for skeleton-based hand gesture recognition. Signal, Image and Video Processing, 17(5): 2521-2529 [DOI: 10.21203/RS3.RS-2285220]

Li Y, He Z H, Ye X, He Z G and Han K R. 2019. Spatial temporal graph convolutional networks for skeleton-based dynamic hand gesture recognition. EURASIP Journal on Image and Video Processing, 2019(1): 1-7 [DOI: 10.1186/S13640-019-0476-X]

Li Y K, Ma D Y, Yu Y H, Wei G S and Zhou Y F. 2021. Compact joints encoding for skeleton-based dynamic hand gesture recognition. Computers and Graphics, 97: 191-199 [DOI: 10.1016/J.CAG.2021.04.017]

Liu J B, Liu Y C, Wang Y, Prinet V, Xiang S M and Pan C H. 2020. Decoupled representation learning for skeleton-based gesture recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5750-5759 [DOI: 10.1109/CVPR42600.2020.00579]

Mai G C, Janowicz K, Yan B, Zhu R, Cai L and Lao N. 2020. Multi-scale representation learning for spatial feature distributions using grid cells [EB/OL]. [2023-06-05]. https://arxiv.org/pdf/2003.00824.pdf

Miah A S M, Hasan M A M and Shin J. 2023. Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model. IEEE Access, 11: 4703-4716 [DOI: 10.1109/ACCESS.2023.3235368]

Molchanov P, Gupta S, Kim K and Kautz J. 2015. Hand gesture recognition with 3D convolutional neural networks//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston, USA: IEEE: 1-7 [DOI: 10.1109/CVPRW.2015.7301342]

Núñez J C, Cabido R, Pantrigo J J, Montemayor A S and Vélez J F. 2018. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition, 76: 80-94 [DOI: 10.1016J.PATCOG.2017.10.033]

Ohn-Bar E and Trivedi M M. 2013. Joint angles similarities and HOG2 for action recognition//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Portland, USA: IEEE: 465-470 [DOI: 10.1109/CVPRW.2013.76]

Oreifej O and Liu Z C. 2013. HON4D: histogram of oriented 4D normals for activity recognition from depth sequences//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 716-723 [DOI: 10.1109/CVPR.2013.98]

Rautaray S S and Agrawal A. 2015. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelli-

gence Review, 43(1): 1-54 [DOI: 10.1007/s10462-012-9356-9]

Shi L, Zhang Y F, Cheng J and Lu H Q. 2021. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer: 38-53 [DOI: 10.1007/978-3-030-69541-5_3]

Shiri F M, Perumal T, Mustapha N and Mohamed R. 2023. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU [EB/OL]. [2023-06-05]. https://arxiv.org/pdf/2305.17473.pdf

Si C Y, Chen W T, Wang W, Wang L and Tan T N. 2019. An attention enhanced graph convolutional LSTM network for skeleton-based action recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1227-1236 [DOI: 10.1109/CVPR.2019.00132]

Song J H, Kong K and Kang S J. 2022. Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network. IEEE Transactions on Circuits and Systems for Video Technology, 32(9): 6227-6239 [DOI: 10.1109/TCSVT.2022.3165069]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang C, Liu Z and Chan S C. 2015. Superpixel-based hand gesture recognition with kinect depth camera. IEEE Transactions on Multimedia, 17(1): 29-39 [DOI: 10.1109/TMM.2014.2374357]

Yan S J, Xiong Y J and Lin D H. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition//Proceedings of the 32nd AAAI Conference on Artificial Intelligence and the 13th Innovative Applications of Artificial Intelligence Conference and 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans, USA: AAAI Press: 7444-7452

作者简介

邓淦森，男，硕士研究生，主要研究方向为深度学习和行为识别。E-mail: denggansen@163.com

丁文文，通信作者，女，副教授，主要研究方向为模式识别。E-mail: dww2048@163.com

杨超，男，硕士研究生，主要研究方向为深度学习和行为识别。E-mail: 1129077634@qq.com

丁重阳，男，博士，主要研究方向为模式识别。E-mail: 281075242@qq.com