(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 113221663 A
(43)申请公布日 2021.08.06

(21)申请号 202110410036.7

(22)申请日 2021.04.16

(71)申请人 南京邮电大学
地址 210023 江苏省南京市栖霞区文苑路9号

(72)发明人 徐小龙 梁吴艳 肖甫

(74)专利代理机构 南京纵横知识产权代理有限公司 32224
代理人 俞翠华

(51)Int.Cl.
G06K 9/00(2006.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

权利要求书3页 说明书11页 附图2页

(54)发明名称

一种实时手语智能识别方法、装置及系统

(57)摘要

本发明公开了一种实时手语智能识别方法、装置及系统，所述方法包括获取手语关节数据和手语骨骼数据；对所述手语关节数据和手语骨骼数据进行数据融合，形成手语关节-骨骼数据；将所述手语关节-骨骼数据分成训练数据和测试数据；获取时空注意力的图卷积神经网络模型，并利用所述训练数据训练所述时空注意力的图卷积神经网络模型，获得训练好的时空注意力的图卷积神经网络模型；将所述测试数据输入至训练好的时空注意力的图卷积神经网络模型，输出手语分类的结果。本发明能够提供一种实时手语智能识别方法，通过从动态骨架数据(手语关节数据和手语骨骼数据)中自动学习空间和时间模式，避免了传统的骨架建模方法对骨架数据建模表达能力有限的问题。

CN 113221663 A

1.一种实时手语智能识别方法,其特征在于,包括:

获取动态骨架数据,所述动态骨架数据包括手语关节数据和手语骨骼数据;

对所述手语关节数据和手语骨骼数据进行数据融合,形成融合后的动态骨架数据,即手语关节-骨骼数据;

将所述手语关节-骨骼数据分成训练数据和测试数据;

获取时空注意力的图卷积神经网络模型,并利用所述训练数据训练所述时空注意力的图卷积神经网络模型,获得训练好的时空注意力的图卷积神经网络模型;

将所述测试数据输入至训练好的时空注意力的图卷积神经网络模型,输出手语分类的结果,完成实时手语智能识别。

2.根据权利要求1所述的一种实时手语智能识别方法,其特征在于,所述手语关节数据的获取方法包括:

利用openpose环境对手语视频数据进行人体关节点2D坐标估计,得到原始关节点坐标数据;

从所述原始关节点坐标数据中筛选出与手语本身的特征直接相关的关节点坐标数据,形成手语关节数据。

3.根据权利要求1或2所述的一种实时手语智能识别方法,其特征在于,所述手语骨骼数据的获取方法包括:

对所述手语关节数据进行向量坐标变换处理,形成手语骨骼数据,每个手语骨骼数据均由源关节和目标关节组成的2维向量表示,每个手语骨骼数据均包含源关节和目标关节之间的长度和方向信息。

4.根据权利要求1所述的一种实时手语智能识别方法,其特征在于:所述手语关节-骨骼数据的计算公式为:

$$\chi_{joints-bonts} = \chi_{joints} \circledast \chi_{bones}$$

其中,⊛代表将手语关节数据和手语骨骼数据在第一维度上连接在一起,x_{joints}、x_{bones}、$x_{joints-bonts}$分别表示手语关节数据、手语骨骼数据以及手语关节-骨骼数据。

5.根据权利要求1所述的一种实时手语智能识别方法,其特征在于:所述时空注意力的图卷积神经网络模型包括顺次相连的归一化层、时空图卷积块层、全局平均池化层和softmax层;所述时空图卷积块层包括顺次设置的9个时空图卷积块。

6.根据权利要求5所述的一种实时手语智能识别方法,其特征在于:所述时空图卷积块包括顺次相连的空间图卷积层、归一化层、ReLU层、时间图卷积层,上一层的输出即为下一层的输入;各每个时空卷积块上均搭建有残差连接。

7.根据权利要求5所述的一种实时手语智能识别方法,其特征在于:设定所述空间图卷积层有L个输出通道和个K个输入通道,则空间图卷积运算公式为:

$$X_{out}^{l} = \sum_{m}^{M} W_{m}^{kl} X_{in}^{k} \left(A_{m}^{r} + Q_{m} + SA_{m} + TA_{m} + STA_{m} \right)$$

其中,X_{out}^{l}表示第L个输出通道的特征向量;X_{in}^{k}表示K个输入通道的特征向量;M表示对一个手语所有节点数的划分方式;W_{m}^{kl}表示在第m子图上的第K行、第L列卷积核;A_{m}^{r}是一个

N×N的邻接矩阵，表示第m子图上的数据节点之间的连接矩阵，r表示利用r阶切比雪夫多项式估计计算捕捉数据节点之间的邻接关系；

Q_m表示一个N×N的自适应权重矩阵，其全部元素初始化为1；

SA_m是一个N×N的空间相关性矩阵，用于确定在空间维度上两个顶点之间是否存在连接以及连接的强度，其表达式为：

$$SA_m = softmax\left(X_{in}^T W_{\theta m}^T W_{\phi m} X_{in}\right)$$

其中，W_θ和W_ϕ分别表示嵌入函数θ(•)和φ(•)的参数；

TA_m是一个N×N的时间相关性矩阵，其元素代表了不同时间段上的节点i和j之间的连接的强弱，其表达式为：

$$TA_m = softmax\left(X_{in}^T W_{\varphi m}^T W_{\psi m} X_{in}\right)$$

其中，W_φ和W_ψ分别表示嵌入函数$\varphi(\cdot)$和ψ(•)的参数；

STA_m是一个N×N的时空相关性矩阵，用于确定时空中两个节点之间的相关性，其表达式为：

$$STA_m = softmax\left(X_{in}^T (W_{\theta m}^T + W_{\varphi m}^T)(W_{\phi m} + W_{\psi m}) X_{in}\right)$$

其中，W_θ和W_ϕ分别表示嵌入函数θ(•)和φ(•)的参数，W_φ和W_ψ分别表示嵌入函数$\varphi(\cdot)$和ψ(•)的参数，X_{in}表示空间图卷积输入的特征向量，X_{in}^T表示对X_{in}转置后的数据。

8.根据权利要求7所述的一种实时手语智能识别方法，其特征在于：所述时间图卷积层属于时间维度的标准卷积层，通过合并相邻时间段上的信息来更新节点的特征信息，从而获得动态骨架数据时间维度的信息特征，各时空卷积块上的卷积操作为：

$$\chi^{(k)} = ReLU\left(\Phi * \left(ReLU\left(\sum_m^M W_m \chi^{(k-1)} \left(A_m^r + Q_m + SA_m + TA_m + STA_m\right)\right)\right)\right)$$

其中，*表示标准卷积运算，Φ为时间维卷积核的参数，其内核大小为K_t×1，ReLU是激活函数，M表示对一个手语所有节点数的划分方式，W_m在第m子图上的卷积核，A_m^r是一个N×N的邻接矩阵，表示第m子图上的数据节点之间的连接矩阵，r表示利用r阶切比雪夫多项式估计计算捕捉数据节点之间的邻接关系，Q_m表示一个N×N的自适应权重矩阵，SA_m是一个N×N的空间相关性矩阵，TA_m是一个N×N的时间相关性矩阵，STA_m是一个N×N的时空相关性矩阵，$x^{(k-1)}$是第k-1的时空卷积块输出的特征向量，$x^{(k)}$汇总了不同时间段中每个手语关节点的特征。

9.一种实时手语智能识别装置，其特征在于，包括：

获取模块，用于获取动态骨架数据，包括手语关节数据和手语骨骼数据；

融合模块，用于对所述手语关节数据和手语骨骼数据进行数据融合，形成融合后的动态骨架数据，即手语关节-骨骼数据；

划分模块，用于将所述手语关节-骨骼数据分成训练数据和测试数据；

训练模块，用于获取时空注意力的图卷积神经网络模型，并利用所述训练数据训练所述时空注意力的图卷积神经网络模型，获得训练好的时空注意力的图卷积神经网络模型；

识别模块，用于将所述测试数据输入至训练好的时空注意力的图卷积神经网络模型，输出手语分类的结果，完成实时手语智能识别。

10.一种实时手语智能识别系统，其特征在于，包括：存储介质和处理器；

所述存储介质用于存储指令；

所述处理器用于根据所述指令进行操作以执行权利要求1-8中任一项所述方法的步骤。

一种实时手语智能识别方法、装置及系统

技术领域

[0001] 本发明属于手语识别技术领域，具体涉及一种实时手语智能识别方法、装置及系统。

背景技术

[0002] 在全球范围内，大约有4.66亿听力受损的人，而且据估计，到2050年该数字高达9亿。手语是一种重要的人类肢体语言表达方式，包含信息量多，同时也是聋哑人与键听人之间沟通的主要载体。因此，利用新兴信息技术对手语进行识别有助于聋哑人与键听人进行实时的交流和沟通，对于改善听障人群的沟通及社交以及促进社会进步具有重要的 现实意义。同时，作为人类身体最直观的表达，手语的应用有助于人机交互向更加自然、便捷的方式升级。因此，手语识别是当今人工智能领域的研究热点。

[0003] 目前，RGB视频和不同类型的模态(例如深度，光流和人体骨骼)都可以用于手语识别(Sign Language Recognition，SLR)任务。与其它模式数据相比，人体的骨架数据不仅能够对人体各个关节之间的关系进行建模和编码，而且对相机拍摄的视角，运动速度，人体外观以及人体尺度等变化具有不变性。更重要的是，它还能够在较高视频帧率下进行计算，这极大的促进在线和实时应用的发展。从历史沿革上，SLR可分为传统识别方法和基 于深度学习的研究方法两大类。2016年以前，基于视觉的传统SLR技术研究较为广泛。传统方法能够解决一定规模下的SLR问题，但算法复杂、泛化性不高，且面向的数据量与 模式种类受限，无法将人类对于手语的智能理解完全表述，例如MEI，HOF以及BHOF等 方法。因此，在当前大数据飞速发展的时代背景下，基于深度学习、挖掘人类视觉与认知 规律的SLR技术成为了必然。当前，大多数已存在的深度学习的研究主要集中在卷积神经 网络(Convolutional Neural Networks，CNN)，循环神经网络(Recurrent Neural Networks，RNN和图卷积网络(Graph Convolutional Networks，GCN)。CNN和RNN非常适合处理 欧几里得数据，例如RGB，depth，光流等，但是，对于高度非线性和复杂多变的骨架数 据却不能很好的表达。GCN很适合处理骨架数据，但是，这种方法在处理面向基于骨架的 手语识别任务时存在以下几个难点：一是仅是利用骨架的关节坐标对手势运动信息进行表 征，这对手和手指运动信息的特征描述还是不够丰富的；二是手语的骨架数据常表现出高 度的非线性和复杂的变化性，这对GCN的识别能力提出了更高的要求；三是主流的基于 骨架的SLR图卷积网络(GCN)倾向于采用一阶Chebyshev多项式近似以减少开销，没有 考虑高阶连接，导致其表征能力受到限制。更糟糕的是，这种GCN网络还缺乏对骨架数 据动态时空相关性的建模能力，无法得到满意的识别精度。

发明内容

[0004] 针对上述问题，本发明提出一种实时手语智能识别方法、装置及系统，通过构建时空 注意力的图卷积神经网络模型，从动态骨架数据(手语关节数据和手语骨骼数据)中自动 学习空间和时间模式，不仅具有更强的表现力，而且具有更强的泛化能力。

[0005] 为了实现上述技术目的，达到上述技术效果，本发明通过以下技术方案实现：

[0006] 第一方面，本发明提供了一种实时手语智能识别方法，包括：

[0007] 获取动态骨架数据，所述动态骨架数据包括手语关节数据和手语骨骼数据；

[0008] 对所述手语关节数据和手语骨骼数据进行数据融合，形成融合后的动态骨架数据，即 手语关节-骨骼数据；

[0009] 将所述手语关节-骨骼数据分成训练数据和测试数据；

[0010] 获取时空注意力的图卷积神经网络模型，并利用所述训练数据训练所述时空注意力的 图卷积神经网络模型，获得训练好的时空注意力的图卷积神经网络模型；

[0011] 将所述测试数据输入至训练好的时空注意力的图卷积神经网络模型，输出手语分类的 结果，完成实时手语智能识别。

[0012] 可选地，所述手语关节数据的获取方法包括：

[0013] 利用openpose环境对手语视频数据进行人体关节点2D坐标估计，得到原始关节点坐 标数据；

[0014] 从所述原始关节点坐标数据中筛选出与手语本身的特征直接相关的关节点坐标数据，形成手语关节数据。

[0015] 可选地，所述手语骨骼数据的获取方法包括：

[0016] 对所述手语关节数据进行向量坐标变换处理，形成手语骨骼数据，每个手语骨骼数据 均由源关节和目标关节组成的2维向量表示，每个手语骨骼数据均包含源关节和目标关节 之间的长度和方向信息。

[0017] 可选地，所述手语关节-骨骼数据的计算公式为：

[0018] $$\chi_{joints-bonts} = \chi_{joints} \circledast \chi_{bones}$$

[0019] 其中，$\circledast$代表将手语关节数据和手语骨骼数据在第一维度上连接在一起，x_{joints}、x_{bones}、$x_{joints-bonts}$分别表示手语关节数据、手语骨骼数据以及手语关节-骨骼数据。

[0020] 可选地，所述时空注意力的图卷积神经网络模型包括顺次相连的归一化层、时空图卷 积块层、全局平均池化层和softmax层；所述时空图卷积块层包括顺次设置的9个时空图卷 积块。

[0021] 可选地，所述时空图卷积块包括顺次相连的空间图卷积层、归一化层、ReLU层、时间 图卷积层，上一层的输出即为下一层的输入；各每个时空卷积块上均搭建有残差连接。

[0022] 可选地，设定所述空间图卷积层有L个输出通道和个K个输入通道，则空间图卷积运 算公式为：

[0023] $$X_{out}^{l} = \sum_{m}^{M} W_m^{kl} X_{in}^{k} \left(A_m^r + Q_m + SA_m + TA_m + STA_m \right)$$

[0024] 其中，X_{out}^{l}表示第L个输出通道的特征向量；X_{in}^{k}表示K个输入通道的特征向量；M表示对一个手语所有节点数的划分方式；W_m^{kl}表示在第m子图上的第K行、第L列卷积核；

[0025] A_m^r是一个N×N的邻接矩阵，表示第m子图上的数据节点之间的连接矩阵，r表示利用r阶切比雪夫多项式估计计算捕捉数据节点之间的邻接关系；

[0026] Q_m表示一个N×N的自适应权重矩阵，其全部元素初始化为1；

[0027] SA_m是一个N×N的空间相关性矩阵，用于确定在空间维度上两个顶点之间是否存在连 接以及连接的强度，其表达式为：

[0028] $$SA_m = softmax\left(X_{in}^T W_{\theta m}^T W_{\phi m} X_{in}\right)$$

[0029] 其中，W_θ和W_ϕ分别表示嵌入函数θ(•)和φ(•)的参数；

[0030] TA_m是一个N×N的时间相关性矩阵，其元素代表了不同时间段上的节点i和j之间的连 接的强弱，其表达式为：

[0031] $$TA_m = softmax\left(X_{in}^T W_{\varphi m}^T W_{\psi m} X_{in}\right)$$

[0032] 其中，W_φ和W_ψ分别表示嵌入函数$\varphi(\cdot)$和ψ(•)的参数；

[0033] STA_m是一个N×N的时空相关性矩阵，用于确定时空中两个节点之间的相关性，其表 达式为：

[0034] $$STA_m = softmax\left(X_{in}^T (W_{\theta m}^T + W_{\varphi m}^T)(W_{\phi m} + W_{\psi m}) X_{in}\right)$$

[0035] 其中，W_θ和W_ϕ分别表示嵌入函数θ(•)和φ(•)的参数，W_φ和W_ψ分别表示嵌入函数$\varphi(\cdot)$和ψ(•)的参数，X_{in}表示空间图卷积输入的特征向量，X_{in}^T表示对X_{in}转置后的数据。

[0036] 可选地，所述时间图卷积层属于时间维度的标准卷积层，通过合并相邻时间段上的信 息来更新节点的特征信息，从而获得动态骨架数据时间维度的信息特征，各时空卷积块上 的卷积操作为：

[0037] $$\chi^{(k)} = ReLU\left(\Phi * \left(ReLU\left(\sum_m^M W_m \chi^{(k-1)}\left(A_m^r + Q_m + SA_m + TA_m + STA_m\right)\right)\right)\right)$$

[0038] 其中，*表示标准卷积运算，Φ为时间维卷积核的参数，其内核大小为$K_t \times 1$，ReLU是 激活函数，M表示对一个手语所有节点数的划分方式，W_m在第m子图上的卷积核，A_m^r是一个N×N的邻接矩阵，表示第m子图上的数据节点之间的连接矩阵，r表示利用r阶切比雪 夫多项式估计计算捕捉数据节点之间的邻接关系，Q_m表示一个N×N的自适应权重矩阵， SA_m是一个N×N的空间相关性矩阵，TA_m是一个N×N的时间相关性矩阵，STA_m是一个 N×N的时空相关性矩阵，$x^{(k-1)}$是第k-1的时空卷积块输出的特征向量，$x^{(k)}$汇总了不同时间 段中每个手语关节点的特征。

[0039] 第二方面，本发明提供了一种实时手语智能识别装置，包括：

[0040] 获取模块，用于获取动态骨架数据，包括手语关节数据和手语骨骼数据；

[0041] 融合模块，用于对所述手语关节数据和手语骨骼数据进行数据融合，形成融合后的动 态骨架数据，即手语关节-骨骼数据；

[0042] 划分模块，用于将所述手语关节-骨骼数据分成训练数据和测试数据；

[0043] 训练模块，用于获取时空注意力的图卷积神经网络模型，并利用所述训练数据训练所 述时空注意力的图卷积神经网络模型，获得训练好的时空注意力的图卷积神经网络模型；

[0044] 识别模块，用于将所述测试数据输入至训练好的时空注意力的图卷积神经网络模型， 输出手语分类的结果，完成实时手语智能识别。

[0045] 第三方面,本发明提供了一种实时手语智能识别系统,包括:存储介质和处理器;

[0046] 所述存储介质用于存储指令;

[0047] 所述处理器用于根据所述指令进行操作以执行第一方面中任一项所述方法的步骤。

[0048] 与现有技术相比,本发明的有益效果:

[0049] (1)本发明借助深层架构强大的端到端自主学习能力来取代传统的人工特征提取:通过构建时空注意力的图卷积神经网络,从动态骨架数据(例如,关节点坐标数据(joints) 和骨骼坐标数据(bones))中自动学习空间和时间模式,避免了传统的骨架建模方法对骨架 数据建模表达能力有限的问题。

[0050] (2)本发明通过利用合适的高阶近似Chebyshev多项式,避免过高的计算开销,同时 扩大GCN的感受野。

[0051] (3)本发明通过设计了一种新的基于注意力的图卷积层,包括空间注意力用于关注 感兴趣的区域,时间注意力用于关注重要的运动信息,以及时空注意力机制用于关注重要 的骨架时空信息,进而实现对重要骨架信息进行选择。

[0052] (4)本发明利用了一种有效的融合策略用于joints和bones数据的连接,不仅避免了 采用双流网络的融合方法带来的内存增加和计算开销,而且能够保证这两种数据的特征在 后期是具有相同维度的。

附图说明

[0053] 为了使本发明的内容更容易被清楚地理解,下面根据具体实施例并结合附图,对本发 明作进一步详细的说明,其中:

[0054] 图1是本发明一种低开销的实时手语智能识别方法的流程图;

[0055] 图2是本发明一种低开销的实时手语智能识别方法中与手语本身直接相关的28个关 节点示意图;

[0056] 图3是本发明一种低开销的实时手语智能识别方法所用的图卷积神经网络模型示意图;

[0057] 图4是本发明一种低开销的实时手语智能识别方法中时空图卷积块示意图;

[0058] 图5是本发明一种低开销的实时手语智能识别方法中时空图卷积示意图;

[0059] 图6是本发明一种低开销的实时手语智能识别方法中时空注意力图卷积层Sgcn示意图;

[0060] 其中,$\odot$代表向量按第一维度连接,$\oplus$是按元素求和,$\otimes$表示矩阵乘法,$\oplus$是按元素求 和。

具体实施方式

[0061] 为了使本发明的目的、技术方案及优点更加清楚明白,以下结合实施例,对本发明进 行进一步详细说明。应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于 限定本发明的保护范围。

[0062] 下面结合附图对本发明的应用原理作详细的描述。

[0063] 实施例1

[0064] 本发明实施例中提供了一种低开销的实时手语智能识别方法，如图1所示，具体包括 以下步骤：

[0065] 步骤1：基于手语视频数据进行骨架数据的获取，包括手语关节数据和手语骨骼数据，具体步骤如下：

[0066] 步骤1.1：搭建openpose环境，包括下载openpose，安装CmakeGui，测试是否安装成功。

[0067] 步骤1.2：利用步骤1.1搭建的openpose环境对手语RGB视频数据进行人体关节点2D 坐标估计，得到130个关节点坐标数据。这里的130个关节点坐标数据包括70个面部关节点，42个手部关节点(左手和右手分别为21个)和18个身体关节点。

[0068] 步骤1.3：利用步骤1.2评估的130个关节点坐标数据，筛选与手语本身的特征直接相 关的关节点坐标数据，作为手语关节数据。对手语本身来说，最直接相关的关节点坐标数 据包括头(1个节点数据)，脖子(1个节点数据)，肩膀(左和右各1个节点数据)， 手臂(左和右各1个节点数据)以及手部(左手和右手各11个节点数据)，共计28个关 节点坐标数据，如图2所示。

[0069] 步骤1.4：将步骤1.3获取的28个关节点坐标数据划分为两个子数据集，即训练数据 和测试数据。考虑到手语样本规模小，在此过程中，利用了3折交叉验证原理，分配80％的样本用于训练和20％的样本用于测试。

[0070] 步骤1.5：对步骤1.4中获得的训练数据和测试数据，分别均进行数据规范化和序列化 处理，从而均生成两个物理文件，用于满足时空注意力的图卷积神经网络模型所需的文件 格式。

[0071] 步骤1.6：利用步骤1.5获取的两个物理文件中的手语关节点数据(joints)进行向量坐 标变换处理，形成手语骨骼数据(bones)，作为新的数据，用于训练和测试，进一步提升 模型的识别率。在这里中，每个手语骨骼数据是通过由两个关节(源关节和目标关节)组成的2维向量表示，其中，源关节点相比目标关节点更靠近骨骼重心。所以说，从源关节 点指向目标关节点的每个骨骼坐标数据是包含两个关节点之间的长度和方向信息的。

[0072] 步骤2，利用数据融合算法实现对步骤1所构建的手语关节数据和手语骨骼数据的数 据融合，形成融合后的融合后的动态骨架数据，即手语关节-骨骼相关数据(joints-bones)。 在数据融合算法中，每个骨骼数据是通过由两个关节(源关节和目标关节)组成的三维向 量表示。考虑到，手语关节数据和手语骨骼数据都是来自相同的视频源，对手语特征描述 的方式是相同的。因此，采用直接将这两种数据在前期输入阶段直接进行融合，不仅可以 保证这两种数据的特征在后期是具有相同维度的。此外，这种前期融合方式还可以避免因 采用双列网络架构进行后期特征融合带来的内存和计算量的增加，如图3所示。具体实现 如下：

[0073] $\chi_{joints-bonts} = \chi_{joints} \circledast \chi_{bones}$

[0074] 其中，$\circledast$代表将手语关节数据和手语骨骼数据在第一维度上连接在一起，x_{joints}、x_{bones}、$x_{joints-bonts}$分别是手语关节数据、手语骨骼数据、手语关节-骨骼数据。

[0075] 步骤3，获取基于时空注意力的图卷积神经网络模型，如图3所示，包括1个归一化层(BN)，9个时空图卷积块(D1-D9)，1个全局平均池化层(GPA)和1个softmax层。

[0076] 按照信息处理顺序依次为：归一化层、时空图卷积块1、时空图卷积块2、时空图卷

积块3、时空图卷积块4、时空图卷积块5、时空图卷积块6、时空图卷积块7、时空图卷 积块8、时空图卷积块9、全局平均池化层、softmax层。其中，9个时空图卷积块的输出 通道参数分别设置为：64，64，64，128，128，128，256，256和256。对于每个时空图 卷积块，均包括空间图卷积层(Sgcn)1、归一化层1、ReLU层1、时间图卷积层(Tgcn) 1；上一层的输出即为下一层的输入；另外，在每个时空卷积块上会搭建残差连接，如图4 所示。每个时空卷积块的空间图卷积层(Sgcn)：对输入骨架数据，即手语关节-骨骼相关 数据(joints-bones)，采用卷积模板在六个通道(Conv-s、Conv-t、)上，对骨架数据进行 卷积操作，得到特征图向量。假定空间图卷积层(Sgcn)有L个输出通道和个K个输入通 道，于是需要利用KL卷积操作实现通道数目的转换，则空间图卷积运算公式为：

[0077] $$X_{out}^{l} = \sum_{m}^{M} W_{m}^{kl} X_{in}^{k} \left(A_{m}^{r} + Q_{m} + SA_{m} + TA_{m} + STA_{m} \right) \quad (1)$$

[0078] 其中，X_{in}^{k}表示K个输入通道的特征向量；X_{out}^{l}表示第L个输出通道的特征向量；M表示 对一个手语所有节点数的划分方式，在这里将一个手语骨架图的邻接矩阵分为三个子图， 即M＝3，如图5(a)空间图卷积所示，不同的颜色深浅的节点代表不同的子图；W_{m}^{kl}表示在第m子图上的第K行、第L列二维卷积核；A_{m}^{r}表示第m子图上的数据节点之间的连接矩阵，r表示利用r阶切比雪夫多项式估计计算捕捉数据节点之间的邻接关系。在这里使用r＝2阶的多项式估 计近似计算公式为：

[0079] $$A_{m}^{2} = 4A^{2} - A - 2I_{n} \quad (2)$$

[0080] 在公式(2)中，A表示一个N×N的邻接矩阵，代表人体自然连接的骨架结构图，I_n是其的单位矩阵，当r＝1时，A_{m}^{r}为邻接矩阵A和单位矩阵I_n的和；Q_m表示一个N×N的自适 应权重矩阵，其全部元素初始化为1；

[0081] SA_m是一个N×N的空间相关性矩阵，用于确定在空间维度上两个节点v_i、v_j之间是否 存在连接以及连接的强度，用normalized embedded Gaussian方程来衡量空间中两个节点 之间的相关性：

[0082] $$f(v_i, v_j) = \frac{exp\left(\theta(v_i)^{T}\phi(v_j)\right)}{\sum_{j=1}^{N} exp\left(\theta(v_i)^{T}\phi(v_j)\right)} \quad (3)$$

[0083] 对于输入的特征图X_{in}^{k}大小为K×T×N，首先用两个嵌入函数θ(•)、φ(•)将其嵌入成 E×T×N，并将其resize成N×ET和KT×N(即改变矩阵的大小)，然后将生成的两个 矩阵相乘得到N×N的相关矩阵SA_m，SA_{m}^{ij}表示节点v_i和节点v_j之间的相关性，因为 normalized Gaussian和softmax操作是等价的，所以公式(3)等同与公式(4)：

[0084] $$SA_{m} = softmax\left(X_{in}^{T} W_{\theta m}^{T} W_{\emptyset m} X_{in}\right) \quad (4)$$

[0085] 其中，W_{θ}和$W_{\emptyset}$分别指嵌入函数θ(•)和φ(•)的参数，在图6中统一被命名cons_s；TA_m是 一个N×N的时间相关性矩阵，用于确定在时间维度上两个节点v_i、v_j之间是否存在连接以 及连接的强度，用normalized embedded Gaussian方程来衡量空间中两个节点之间的相关 性：

[0086] $$f(v_i,\ v_j)=\frac{exp\left(\varphi(v_i)^T\psi(v_j)\right)}{\sum_{j=1}^{N}exp\left(\varphi(v_i)^T\psi(v_j)\right)} \quad (5)$$

[0087] 对于输入的特征图X_{in}^{k}大小为K×T×N,首先用两个嵌入函数$\varphi(\cdot)$、ψ(•)将其嵌入成 E×T×N,并将其resize成N×ET和KT×N,然后将生成的两个矩阵相乘得到N×N的 相关矩阵TA_m,TA_m^{ij}表示节点v_i和节点v_j之间的时间相关性,因为normalized、Gaussian 和softmax操作是等价的,所以公式(5)等同与公式(6):

[0088] $$TA_m=softmax\left(X_{in}^{T}W_{\varphi m}^{T}W_{\psi m}X_{in}\right) \quad (6)$$

[0089] 其中,W_φ和W_ψ分别指嵌入函数$\varphi(\cdot)$和ψ(•)的参数,在图6中统一被命名cons_t;STA_m是一个N×N的时空相关性矩阵,用于确定在时空维度上两个节点v_i、v_j之间是否存在连接 以及连接的强度,使用空间SA_m和时间TA_m这两个模块直接构建,用于确定时空中两个节点之间的相关性,对于输入的特征图X_{in}^{k}大小为K×T×N,首先用四个嵌入函数θ(•)、Φ(•)、 $\varphi(\cdot)$、ψ(•)将其嵌入成E×T×N,并将其resize成N×ET和KT×N,然后将生成的两个矩阵相乘得到M×N的相关矩阵STA_m,STA_m^{ij}表示节点v_i和节点v_j之间的时空相关性,并由空间SA_m和时间TA_m这两个模块直接构建:

[0090] $$STA_m=softmax\left(X_{in}^{T}\left(W_{\theta m}^{T}+W_{\varphi m}^{T}\right)\left(W_{\phi m}+W_{\psi m}\right)X_{in}\right) \quad (7)$$

[0091] 其中,W_θ和W_ϕ分别指嵌入函数θ(•)和Φ(•)的参数,在图6中统一被命名cons_s,W_φ和W_ψ分 别指嵌入函数$\varphi(\cdot)$和ψ(•)的参数,在图6中统一被命名cons_t。

[0092] 每个时空卷积块的时间图卷积Tgcn层:在时间图卷积Tgcn中采用时间维度的标准卷积对 得到特征图通过合并相邻时间段上的信息来更新节点的特征信息,从而获得节点数据时间 维度的信息特征,如图5(b)时间图卷积所示,以第k个时空卷积块上卷积操作为例:

[0093] $$\chi^{(k)}=ReLU\left(\Phi*\left(ReLU\left(\sum_m^M W_m X^{(k-1)}\left(A_m^r+Q_m+SA_m+TA_m+STA_m\right)\right)\right)\right) \quad (8)$$

[0094] 其中*表示标准卷积运算,Φ为时间维卷积核的参数,其内核大小为K_t×1,在这里取 K_t=9,激活函数是ReLU,M表示对一个手语所有节点数的划分方式,W_m在第m子图上的 卷积核,A_m^r是一个N×N的邻接矩阵,表示第m子图上的数据节点之间的连接矩阵,r表 示利用r阶切比雪夫多项式估计计算捕捉数据节点之间的邻接关系,Q_m表示一个N×N的 自适应权重矩阵,SA_m是一个N×N的空间相关性矩阵,TA_m是一个N×N的时间相关性矩阵, STA_m是一个N×N的时空相关性矩阵,$x^{(k-1)}$是第k-1个时空卷积块输出的特征向量,$x^{(k)}$汇 总了不同时间段中每个手语关节点的特征。

[0095] ReLU层:在ReLU层中采用线性整流函数(RectifiedLinear Unit,ReLU)对得到的特征 向量,线性整流函数为:Φ(x)=max(0,x)。其中x为ReLU层的输入向量,X(x)为输出向量,作为下一层的输入。ReLU层能更加有效率的梯度下降以及反向传播,避免了梯度爆炸和梯度消失问题。同时ReLU层简化了计算过程,没有了其他复杂激活函数中诸如指数函数的影响;同时活跃度的分散性使得卷积神经网络整体计算成本下降。在每个图卷积操作之

后，都有ReLU的附加操作，其目的是在图卷积中加入非线性，因为使用图卷积来解决的现实世界的问题都是非线性的，而卷积运算是线性运算，所以必须使用一个如ReLU的激活函数来加入非线性的性质。

[0096] 归一化层(BN)：归一化有助于快速收敛；对局部神经元的活动创建竞争机制，使得其中响应比较大的值变得相对更大，并抑制其他反馈较小的神经元，增强了模型的泛化能力。

[0097] 全局平均池化层(GPA)：对输入的特征图进行压缩，一方面使特征图变小，简化网络 计算复杂度；一方面进行特征压缩，提取主要特征。全局平均池化层(GPA)可以在保持最重要的信息的同时降低特征图的维度。

[0098] 步骤4，利用所述训练数据对时空注意力的图卷积神经网络模型进行训练，具体步骤 如下：

[0099] 步骤4.1，随机初始化所有时空注意力的图卷积神经网络模型的参数和权重值；

[0100] 步骤4.2，将融合的动态骨架数据(手语关节-骨骼数据)作为时空注意力的图卷积网 络模型该模型的输入，经过前向传播步骤，即归一化层、9个时空图卷积块层、全局平均池化层，最后达到softmax层进行分类，得到分类结果，也就是输出一个包含每个类预测的概率值的向量。由于权重是随机分配给第一个训练样例的，因此输出概率也是随机的；

[0101] 步骤4.3，计算输出层(softmax层)的损失函数Loss，如式(9)所示，采用交叉熵(Cross Entropy)损失函数，定义如下：

[0102] $$Loss = -\frac{1}{n}\sum_{1}^{n}\sum_{k}^{C} y_k \log P_k \qquad (9)$$

[0103] 其中，C是手语分类的类别数，n是样本的总数量，x_k是softmax输出层第k个神经元的输出，P_k是模型预测的概率分布，即softmax分类器对每个输入的手语样本属于第K个类别的概率 计算，y_k是真实手语类别的离散分布。Loss表示损失函数，用来评测模型对真实概率分布 估计的准确程度，可以通过最小化损失函数Loss来优化模型，更新所有的网络参数。

[0104] 步骤4.4，使用反向传播计算网络中所有权重的误差梯度。并使用梯度下降更新所有 滤波器值、权重和参数值，以最大限度地减少输出损失，也就是损失函数的值尽量降低。权重根据它们对损失的贡献进行调整。当再次输入相同的骨架数据时，输出概率可能更接近目标矢量。这意味着网络已经学会了通过调整其权重和滤波器来正确分类该特定骨架，从而减少输出损失。滤波器数量，滤波器大小，网络结构等参数在步骤4.1之前都已经固定，并且在训练过程中不会改变，只更新滤波器矩阵和连接权值。

[0105] 步骤4.5，对训练集中的所有骨架数据重复步骤4.2-4.4，直到训练次数达到设定的 epoch值。完成上述步骤对训练集数据通过构建的时空注意力的卷积神经网络进行训练学 习，这实际上意味着GCN的所有权重和参数都已经过优化，可正确手语分类。

[0106] 步骤5，用已经训练完成的时空注意力的图卷积神经网络模型对测试样本进行识别，并输出手语分类的结果。

[0107] 根据输出手语分类的结果统计出识别的准确率。其中以识别精度(Accuracy)作为评 价系统的主要指标，包括Top1和Top5精确度，它的计算方式为：

[0108] $$Accuracy = \frac{(TP + TN)}{(P + N)}$$

[0109] 其中TP为被正确地划分为正例的个数,即实际为正例并且被分类器划分为正例的实例 数;TN为被正确地划分为负例的个数,即实际为负例并且被分类器划分为负例的实例数; P为正样本数,N为负样本数。通常来说,准确率越高,识别结果越好。在这里,假设分类类别共n类,现在有m测试样本,那么一个样本输入网络,得到n类别概率,Top1是这n 类别概率中取概率最高的一个类别,如果此测试样本的类别是概率最高类别,则表明预测 正确,反之预测错误,Top1正确率是预测正确样本数/所有样本数,属于普通的准确率 Accuracy;而Top5就是这n类别概率中取概率最高的前五个类别,如果此测试样本的类别 在这五个类别中,则表明预测正确,反之预测错误,Top5正确率是预测正确样本数/所有 样本数。

[0110] 为了说明数据融合策略和空间图卷积Sgcn的5个模块对时空注意力的图卷积神经网络 模型的有效性,通过在预处理后的DEVISIGN-D手语骨架数据上进行了实验,首先将ST-GCN的模型作为基准模型,而后逐步加入空间图卷积Sgcn的各个模块。表1反映了使 用不同模式数据的时空图卷积神经网络模型的最佳分类能力,在这里,时空注意力的图卷 积神经网络模型,记为model。

[0111] 表1各个模型和融合框架在DEVISIGN-D上的实验结果

[0112]

训练数据源	训练的策略	Top1（%）	Top5（%）
Joints 数据	ST-GCN	**74.44**	**92.79**
	model+Q_m	**79.42**	**94.37**
	model+A_m^r	79.69	94.68
	model+SA_m	79.78	94.70
	model +TA_m	79.79	94.73
	model+STA_m	**79.89**	**94.75**
Bones 数据	ST-GCN	73.90	91.92
	model+Q_m	78.74	94.31
	model+A_m^r	79.28	94.64
	model+SA_m	79.78	94.69
	model+TA_m	79.82	94.73
	model+STA_m	**79.84**	**94.76**
数据融合		**80.73**	**95.41**

[0113] 对比表1中的数据可以发现,在joints数据模式下,使用Q_m相比较基准方法可以对识 别Top1正确率有5.02%以上的提升,这也验证了在考虑给定图中每个节点之间连接一定的 权值参考下,有利于手语的识别。另外,实验结果也表明引入高阶Chebyshev近似A_m^r是可 以使图卷积神经网络的感受野增大,有效提升手语识别的正确率。主要是因为感受野的值 越大,表示其能接触到的原始骨架图的范围就越大,也意味着它可能蕴含了更为全局,语 义层次更高的特征;相反,值越小则表示其所包含的特征越趋向局部和细节。输入的骨架 数据是3D的数据,与2D图像相比多了一个时间维度,与1D语音信号数据相比多了一个空间维度。因此,训练阶段引入空间注意力模块SA_m,时间注意力模块TA_m和时空注意力 模块STA_m,该方法能够很好地关注感兴趣的区域和选择用于关注重要的运动信息。实验结 果表明,该注意力机制的模块能够有效提升手语识别地正确率。同时从表格1可以发现, 当分别以一阶joints数据、二阶bones数据训练模型时,由于一阶关节数据源将人体从复 杂的背

景图像中区分出来,代表了人体骨架的关节数据特征,因此它的识别效果有微弱的 优势。而在融合两种数据后,识别正确率有了进一步的提高,这主要是因为joints数据对 人体骨架的识别效果好,二阶bones数据则更加注重人体骨架中骨骼的细节变化,所以, 当这两种数据融合则会增强模型对不同数据中的运动信息的学习能力。也就是说,这两种 新的数据对手势识别一样有用,并且可以将这两种数据进行前期融合后用于训练模型,能 够达到进一步提升识别精度的效果。

[0114] 为了进一步验证时空注意力的图卷积神经网络模型的优势,本实验将其与公开方法在 识别率Accuracy方面进行对比,如表2所示,在这里时空注意力的图卷积神经网络模型, 记为model。

[0115] 表2本发明方法和其它公开方法在ASLLVD上的识别结果

[0116]

方法	*Accuracy* (%)
MHI	10.00
MEI	25.00
PCA	45.00
ST-GCN	**56.82**
HOF	70.00
BHOF	85.00
Model	**87.88**

[0117] 如表格2所示,以前的研究呈现出更多原始方法,诸如MEI和MHI等方法,它们主要是从连续动作视频帧之间的差异来检测运动及其强度。它们不能区分个体,也不能专注于身体的特定部位,导致任何性质的运动被认为是对等的。而PCA反过来又增加了基于对方差较大的组分的识别来降低组分维数的能力,从而使之与检测更为相关框架内的运动。基于时空图卷积网络(ST-GCN)的方法是利用人体骨架的图结构,重点关注身体的运动及 其各部分之间的相互作用,并忽略周围环境的干扰。此外,在空间和时间维度下的运动, 能够捕捉到随着时间推移所进行的手势动作的动态方面的信息。基于这些特点,该方法很 适用于处理手语识别所面临的问题。相比于模型ST-GCN,时空注意力的图卷积模型(model) 更深入,尤其是对于手和手指的运动。为了寻找能够丰富手语运动的特征描述,model还 将二阶骨骼数据bones用于提取手语骨架的骨骼信息。此外,为了提升图卷积的表征能力, 扩大GCN的感受野,model采用合适的高阶Chebyshev近似的计算。最后,为了进一步改 善GCN的性能,注意力机制用于实现对手语骨架相对重要信息的选择,进一步提升图的 节点进行正确分类。表2的实验结果表明融合了joints和bones这两种数据的model模型 明显优于现有的基于ST-GCN的手语识别方法,正确率提升了31.06%。利用HOF特征提 取技术对图像进行预处理,可以为机器学习算法提供更丰富的信息。而BHOF方法则是应 用连续步骤进行光流提取、彩色地图创建、块分割和从中生成直方图,能够确保提取出更 多关于手运动的增强特征,有利于其符号识别性能。该技术源于HOF,不同之处在于计算 光流直方图时,只关注个体的手部。而像基于ST-GCN时空图卷积网络仅是基于人体关节 的坐标图,还不能提供像BHOF那样显著的结果,但是model的方法可以与BHOF方法相 媲美,正确识别率提升了2.88%。

[0118] 实施例2

[0119] 基于与实施例1相同的发明构思，本发明实施例中提供了一种实时手语智能识别装置，包括：

[0120] 其余部分均与实施例1相同。

[0121] 实施例3

[0122] 基于与实施例1相同的发明构思，本发明实施例中提供了一种实时手语智能识别系统，包括存储介质和处理器；

[0123] 所述存储介质用于存储指令；

[0124] 所述处理器用于根据所述指令进行操作以执行根据实施例1中任一项所述方法的步骤。

[0125] 本发明提出的一种低开销的实时手语智能识别方法不仅通过利用合适的高阶近似扩 大GCN感受野，进一步提升GCN的表征能力，而且采用注意力机制为每个手势动作选择最丰富最重要的信息。其中，空间注意力用于关注感兴趣的区域，时间注意力用于关注重要的运动信息，以及时空注意力机制用于关注重要的骨架时空信息。此外，该方法还从原始的视频样本中提取骨架样本包括joints和bones作为模型的输入，并采用深度学习的前期融合策略对joints和bones数据的特征进行融合。其中这种前期的融合策略不仅避免了采用双流网络的融合方法带来的内存增加和计算开销，而且能够保证这两种数据的特征在后期是具有相同维度的。实验结果显示，该方法在DEVISIGN-D和ASLLVD数据集上的 TOP1分别可达80.73％和87.88％，TOP5分别可达95.41％和100％。这一结果验证了这种方 法进行动态骨架手语识别方法的有效性。总之，在基于聋哑人的手语识别任务中，该方法 具有明显的优势，特别适合用于复杂和多变的手语识别。

[0126] 以上显示和描述了本发明的基本原理和主要特征和本发明的优点。本行业的技术人员 应该了解，本发明不受上述实施例的限制，上述实施例和说明书中描述的只是说明本发明 的原理，在不脱离本发明精神和范围的前提下，本发明还会有各种变化和改进，这些变化 和改进都落入要求保护的本发明范围内。本发明要求保护范围由所附的权利要求书及其等 效物界定。

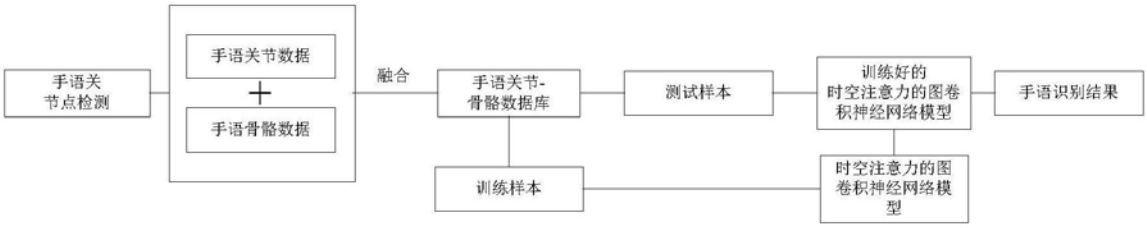
手语关节点检测
手语关节数据
+
手语骨骼数据
融合
手语关节-骨骼数据库
测试样本
训练好的时空注意力的图卷积神经网络模型
手语识别结果
训练样本
时空注意力的图卷积神经网络模型

图1

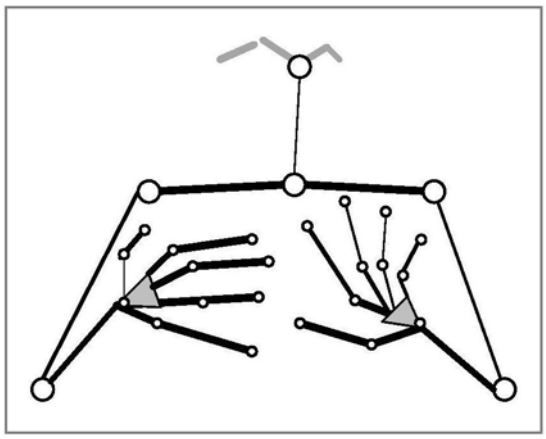

图2

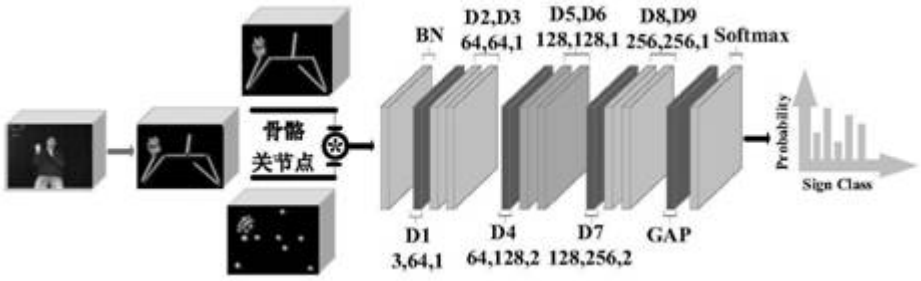
骨骼
关节点
BN
D2,D3
64,64,1
D5,D6
128,128,1
D8,D9
256,256,1
Softmax
Probability
Sign Class
D1
3,64,1
D4
64,128,2
D7
128,256,2
GAP

图3

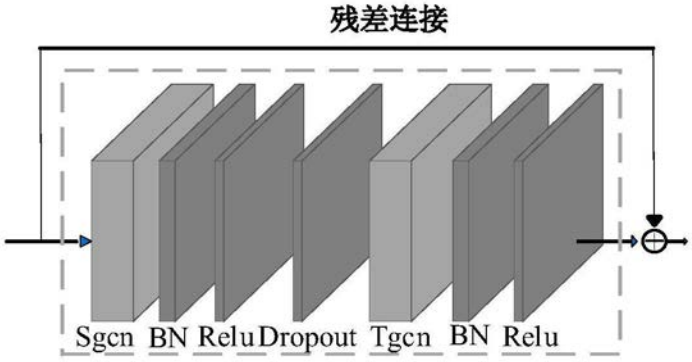
残差连接
Sgcn BN Relu Dropout Tgcn BN Relu

图4

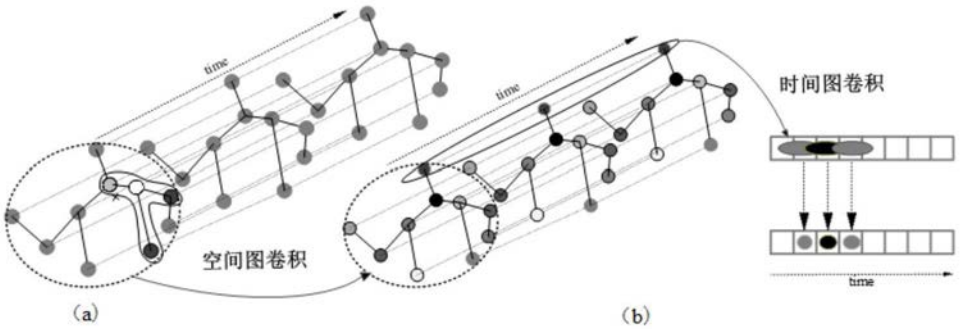
time
time
时间图卷积
空间图卷积
time
(a)
(b)

图5

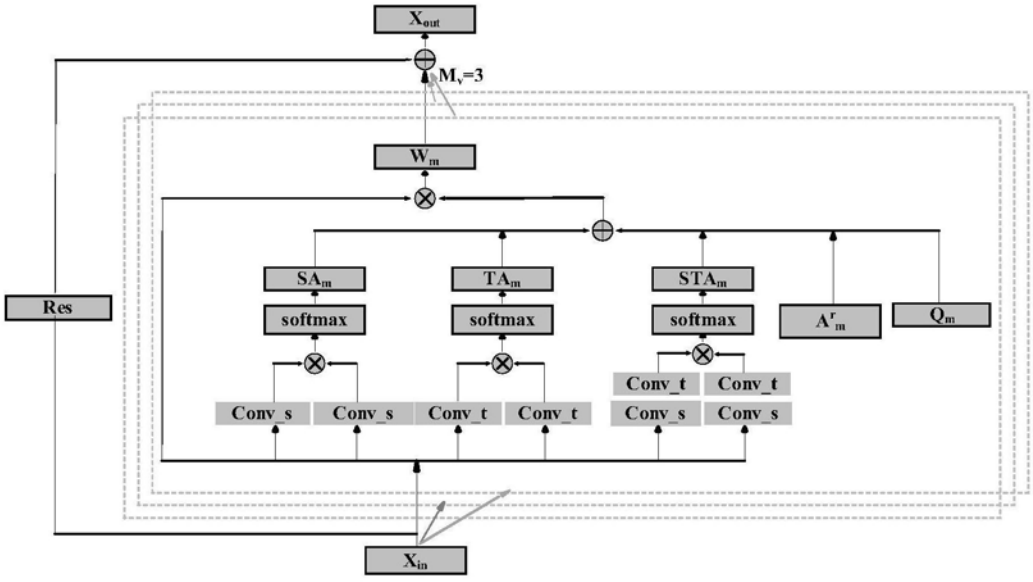
X_{out}
M_v=3
W_m
SA_m
TA_m
STA_m
softmax
softmax
softmax
A^r_m
Q_m
Res
Conv_s
Conv_s
Conv_t
Conv_t
Conv_t
Conv_t
Conv_s
Conv_s
X_{in}

图6