◆交通场景◆

一种基于人体姿态的新型中国交警手势识别网络*

覃　晓[1,2],李永玉[1],吴琨生[3],元昌安[4**],谭思靖[1],刘善锐[1]

(1.南宁师范大学,广西人机交互与智能决策重点实验室,广西南宁　530100;2.广西区域多源数据集成与智能处理协同创新中心,广西桂林 541004;3.广西壮族自治区南宁树木园,广西南宁　530225;4.广西科学院,广西南宁　530012)

摘要:交警手势识别对于自动驾驶技术至关重要,现有的基于人体姿态的交警手势识别方法在骨架特征提取中存在特征不完整、鲁棒性不足等问题;时序特征提取存在动态信息丢失、时序依赖性弱、实时性差等问题,其效果也极易受到环境背景的影响。本研究提出一种基于人体姿态的新型交警手势识别网络(Pose Long Short-Term Memory,PoseLSTM)。PoseLSTM 中的关节组合编码器(Compositional Tokens Multi-layer perceptron Mixer,CTMM)能够捕捉身体各关节间的关联特征,并通过依赖建模来转换这些关节信息,形成多部位特征表示,解决了基于长短期记忆(Long Short-Term Memory,LSTM)的算法无法有效提取骨架特征的问题;此外,PoseLSTM 中的混合架构注意力 LSTM (Attention LSTM),能更好地融合输入与隐藏状态的信息,其效果优于原始 LSTM。实验结果表明,PoseLSTM 在开源的中国交警手势数据集上的准确率为 100.00%,实现了最优。为了证明 PoseLSTM 的泛化能力,在开放手语数据集 LSA64、WLASL-100 和 CSL-500 上进行实验,其准确率分别达到 100.00%、59.69%和 96.40%。

关键词:交警手势识别;注意力机制;LSTM;关节组合

中图分类号:S781,TP391.41,TP183　　**文献标识码**:A　　**文章编号**:1005-9164(2024)05-1011-14

DOI:10.13656/j.cnki.gxkx.20241127.017

2021 年 2 月,中共中央、国务院发布的《国家综合立体交通网规划纲要》[1]将智能驾驶技术列为国家战略的重点发展方向之一。目前,国内多个城市正积极推进无人驾驶出租车的测试工作,这些无人驾驶车辆在实际道路环境上运行时,必须能够准确地理解和遵循交警的指挥手势。因此,提高交警手势的识别效率与准确性对智能驾驶技术至关重要。

交警共有停止、直行、左转等 8 种基本交通指挥手势,这些手势对于维护道路秩序,特别是在交通信号系统无法覆盖或出现故障的情况下尤为重要。交警手势识别属于动作识别,基于骨架的交警手势识别方法更专注于动作的变化和紧凑性,不易受环境、背

收稿日期:2024-07-22　　修回日期:2024-09-25

*科技部科技创新 2030—"脑科学与类脑研究"重大项目(2021ZD0201904)和广西科技重大专项(桂科 AA22068057)资助。

【第一作者简介】

覃　晓(1973—),女,硕士,教授,主要从事数字图像处理和自然语言理解研究,E-mail:7670172@qq.com。

【**通信作者简介】

元昌安(1964—),男,博士,教授,主要从事智能计算研究,E-mail:yuanchangan@126.com。

【引用本文】

覃晓,李永玉,吴琨生,等.一种基于人体姿态的新型中国交警手势识别网络[J].广西科学,2024,31(5):1011-1024.

QIN X,LI Y Y,WU K S,et al. A Novel Chinese Traffic Police Gesture Recognition Network Based on Human Pose [J]. Guangxi Sciences,2024,31(5):1011-1024.

景的干扰,近年来越来越受到关注。其一般流程如下:首先,需要使用训练好的人体姿态估计器检测每个时间步长下的交警位置并提取 N 个人体骨骼关节点坐标;然后,根据提取的坐标信息提取每个时间步长下的人体骨骼特征与时序特征;最后,将提取到的特征输入到分类层进行手势分类。因此现有的大部分方法运用骨骼关节点坐标构建图卷积神经网络(Graph Convolutional Networks,GCN)[2-5],以便更好地提取人体骨架深层特征。GCN 将每个时间步长的每个人体关节视为一个节点,沿着空间维度和时间维度将相邻节点用边连接;然后将图卷积层应用于构建的图,以发现跨空间和时间的动作特征。由于其良好的性能,GCN 一直是处理骨架序列的标准方法,在此类任务中取得了相当大的成功[6,7]。Xiong 等[8]和 Liu 等[9]成功地将时空图卷积网络(Spatial Temporal Graph Convolutional Networks,ST-GCN)应用在中国交警手势识别任务中,并且实现了高准确率。虽然基于 GCN 的方法已经较为先进,但仍存在一些不足:①鲁棒性差,虽然 GCN 可以直接处理人体的骨架坐标信息,但其识别能力受到坐标分布偏移的显著影响,这对姿态估计器的准确性提出了更高的要求,因为坐标中的微小扰动通常会导致完全不同的预测[10]。②泛化能力差,由于 GCN 将每个人体关节视为一个节点,因此 GCN 的复杂度与骨骼关节点的数量成线性关系。在涉及大量关节点的任务中,数据可能变得稀疏,导致模型难以学习到有效的特征表示,限制了其适用于关节点较多的任务。③推理计算消耗大,尽管基于 GCN 的方法在处理骨架数据的拓扑结构上表现优异,但在捕获与整合时间序列动态特征上仍存在局限性。这是因为 ST-GCN 都是并行运算,每次推理需要同时处理整个序列,处理长序列时效率低下。

长短时记忆(Long Short-Term Memory,LSTM)网络是循环神经网络(Recurrent Neural Network,RNN)的一种变体,解决了 RNN 无法建立长期记忆的缺点。交警的手势具有时间上的连贯性和周期性特征,LSTM 能较好地解决交警手势识别的时序特征提取问题[11,12]。但是,基于 LSTM 的方法在交警手势识别中的准确率通常低于基于 GCN 的方法,这不仅因为基于 LSTM 的方法缺乏有效的骨架特征提取模块,LSTM 自身的上下文依赖能力不足也是原因之一[13]。为解决这些问题,本研究基于改进的 LSTM 算法,提出一种基于人体姿态的新型交警手势识别网络(Pose Long Short-Term Memory,PoseLSTM),其设计思路与基于 GCN 的方法完全不同。改进后的 LSTM 不仅保持了快速推理的优点,还提高了交警手势识别的准确率,并解决了基于 GCN 方法的一些局限性。为了有效提取骨架特征,受 Geng 等[14]提出的 Pose as Compositional Tokens(PCT)启发,本研究提出关节组合编码器(Compositional Tokens Multi-layer perceptron Mixer,CTMM),CTMM 将每个时间步的人体骨架信息拆分为 M 个子结构,并通过下采样操作得到骨架特征,代表人体此时的运动状态。此外,为优化实时推理性能并强化信息传播效率,本研究提出注意力 LSTM(Attention LSTM),用来提取时序特征。Attention LSTM 通过引入注意力门控机制,加强原始 LSTM 中当前输入信息和隐藏状态的相互作用并聚焦两者之间的关键信息。Attention LSTM 不需要追溯并重新计算整个序列的历史信息,从而避免了随着序列长度增加而带来的计算负担呈指数级增长的问题,有利于实现快速推理及内存高效利用。交警手势识别过程如图 1 所示,首先将二维(2D)姿态坐标输入到 PoseLSTM 中,并使用 CTMM 提取骨架特征,再将骨架特征输入到 Attention LSTM 中提取时序特征,最后通过全连接层实现手势分类。

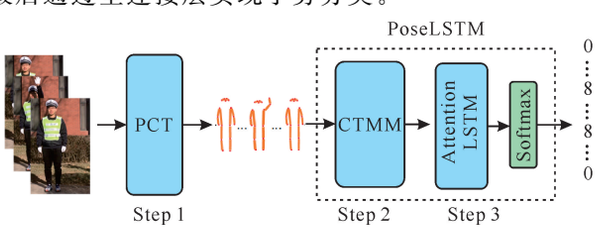

图 1 交警手势识别过程

Fig. 1 Traffic police gesture recognition process

1 相关工作

1.1 基于 2D 人体姿态估计的交警手势识别

目前,2D 人体姿态估计技术的进步显著提升了动作识别方法的性能,尤其是依赖于高精度关节位置信息的方法。Yan 等[7]率先提出一种用于动作识别的 ST-GCN,它能够有效地学习人体骨架的时间动态并进行分类,通过人体骨架特征也能实现高效的姿势识别[15-17]。刘永涛等[2]改进了姿态估计算法,将 ST-GCN 应用到了交警手势识别中,并实现了高准确率的识别。Geng 等[14]提出的 PCT 在处理遮挡问题上取得了较好的性能,成功实现了在遮挡条件下对人体关节的精确提取;PCT 不直接对整个人体进行一次

性处理,而是将其分解为一系列更细粒度的“组成标记”,这些标记代表人体的关键关节或特定部位,通过它们的不同组合和排列来构建完整的姿态;这种方法有助于模型更好地理解人体的不同组成部分,如上肢、下肢和躯干等,即使在部分遮挡的情况下也能保持较好的识别效果;PCT 的快速推理能力和对遮挡情况的有效处理能力使得它能够应用于实时监控和智能决策系统中。PoseLSTM 使用 PCT 所提取的人体 17 个关节来表示一个姿势,如图 2 所示。此外,基于 PCT,本研究还提出了能够有效提取人体各个部位运动特征的 CTMM。

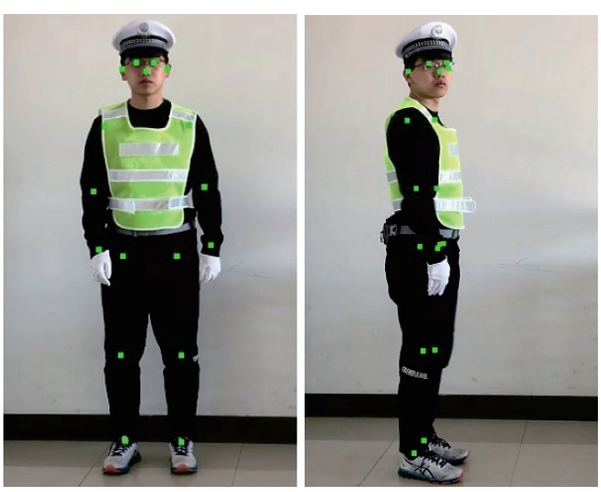

图 2　PCT 在正面(左)和侧面(右)关节点提取情况

Fig. 2　Joint point extraction of PCT on the front (left) and side (right)

1.2　Mogrifier LSTM

在传统的 LSTM 模型中,当前时刻的输入 x_t 和隐藏状态 h_{t-1} 在进行更新前是完全独立的,会导致部分信息在门控计算过程中流失,未能充分利用两者之间的内在关联性[13]。为解决该问题,Mogrifier LSTM[13]在 LSTM 前面添加了一个新的门控机制,Mogrifier LSTM 通过引入矩阵 Q 和 R 以及超参数 r 来增强 x_t 与 h_{t-1} 之间的信息交互,超参数 r 控制 x_t 和 h_{t-1} 应进行交互计算的次数。通过 r 次迭代,将先前的状态和输入信息通过特定的计算方式相互作用,增强了 LSTM 记忆单元对序列中不同时间步信息的捕捉和处理能力,以更好地捕捉序列中的长期依赖关系。Mogrifier LSTM 输入 x_t 与隐藏状态 h_{t-1} 计算公式为

$$x_t^i = 2\sigma(Q^i h_{t-1}^{i-1}) \odot x_t^{i-1}, i \in [1, \cdots, \lfloor (r+1)/2 \rfloor], \tag{1}$$

$$h_{t-1}^j = 2\sigma(R^j x_t^j) \odot h_{t-1}^{j-1}, j \in [1, \cdots, \lfloor r/2 \rfloor], \tag{2}$$

其中,$r \in N$,Q、R 是可训练的权重;σ 是激活函数;$\odot$表示哈达玛积,元素对应相乘;$\lfloor \cdot \rfloor$表示向下取整运算,$r=0$ 时,整个模型退化为原始的 LSTM。但是,Mogrifier LSTM 中的矩阵 Q 和 R 是固定的,与 x_t 和 h_{t-1} 无关,因此当相似手势的前 t 帧运动轨迹相同时,这种计算方式会破坏相似手势之间的差异。本研究在 LSTM 中引入注意力机制替换 Mogrifier LSTM 的门控机制,提出 Attention LSTM,Attention LSTM 通过注意力门控机制增强了模型对序列中重要信息的关注程度,捕捉帧与帧之间的细粒度信息,有助于区分相似手势并保留关键细节,从而在交警手势识别任务中表现出更好的性能。

2　方法

2.1　关节组合编码器(CTMM)

首先,本研究提出一个骨架特征提取模块 CTMM,如图 3 所示,CTMM 考虑了身体运动时各个部位的变化及其被拆分的方式。在手势执行过程中,除了直观可见的头、手、手臂、脚等由相邻关节点组成的部位(红色方框表示)之外,还包括一些隐藏部位。这些隐藏部位可能是由几个不相邻的关节点构成(黄色线段表示),尽管无法直接观察到,但它们的变化同样代表着手势的运动情况。CTMM 的目标是从骨架数据中学习如何组合这些关节点,形成 M 个由不同关节点组成的组合。这一过程相当于将完整的骨架结构分解为M个独立的子结构,每个子结构对应身体

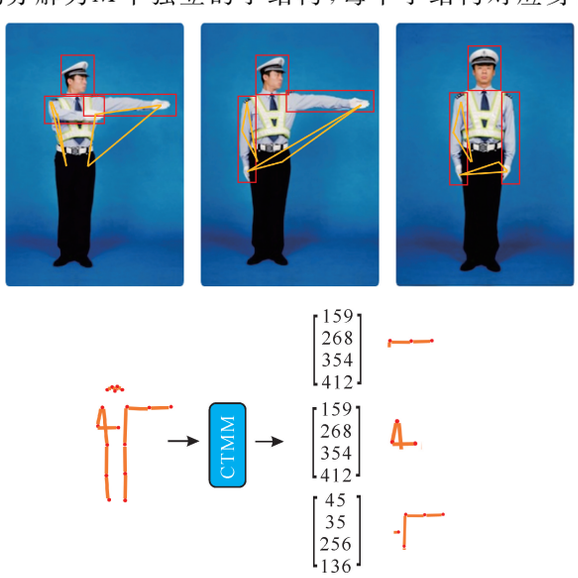

图 3　CTMM 骨架拆分示意图(以直行手势为例)

Fig. 3　Schematic of CTMM split skeleton (take the straight ahead gesture as the example)

的一个特定部位,CTMM 对这些子结构分别进行特征提取,旨在捕捉每个部位特有的动态和形态信息。然后使用全局平均池化将每个部位的特征压缩为一个固定大小的向量,池化策略有助于保留重要信息的同时减少冗余,从而增强模型的泛化能力。最后,这些经过池化的特征向量被拼接起来,形成一个能够全面反映整个骨架姿态信息的完整姿态特征向量。这种方法可以使模型适应于关节点数量不同的交警手势识别任务,提高效率。CTMM 模型结构见图 4,图中 B,L 分别为 batch size 和输入的视频序列长度。CTMM 具体的计算过程如下。

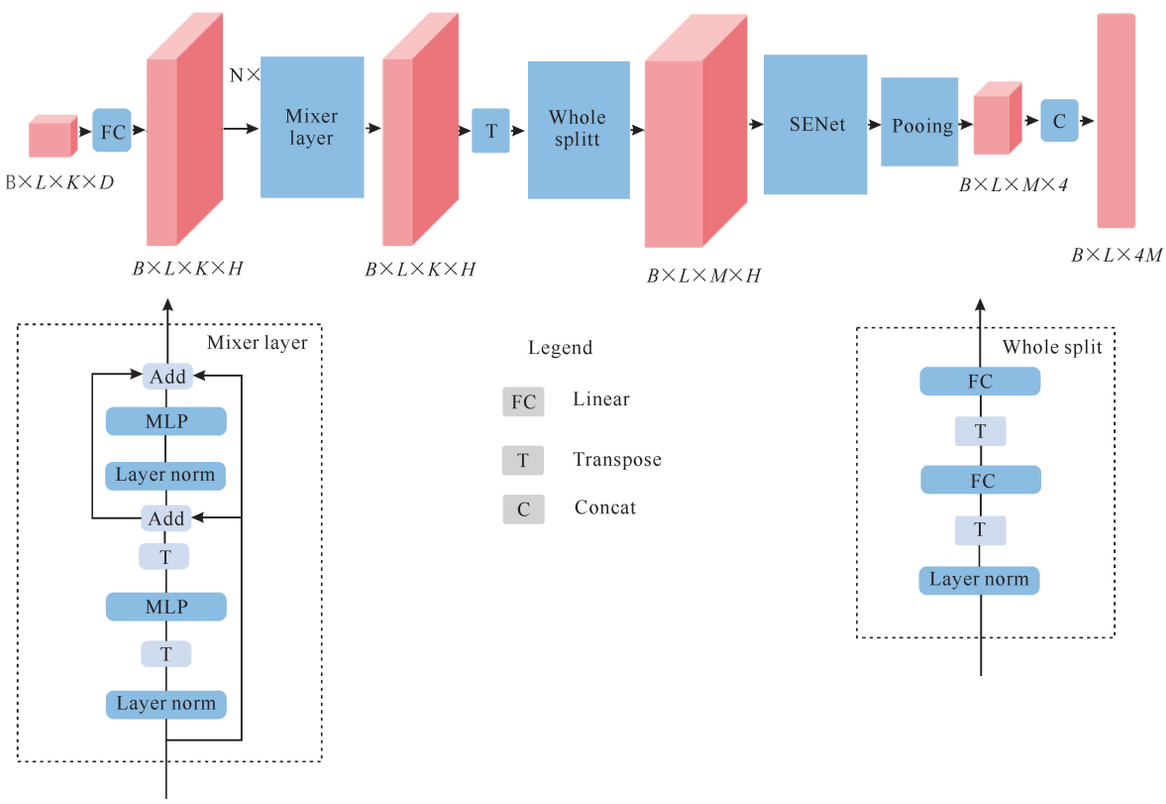

图 4　CTMM 模块结构

Fig. 4　Structure of the CTMM

首先,每一帧的原始姿态被表示为 $G \in R^{K\times D}$,K 是提取的人体关节的数量,D 为坐标的维度,默认为二维坐标,$D=2$。使用一个线性投影层将人体的 K 个关节升维:

$$G=\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_K \end{bmatrix}, \bar{G}=\begin{bmatrix} \bar{g}_1 \\ \bar{g}_2 \\ \vdots \\ \bar{g}_K \end{bmatrix}=GW, \tag{3}$$

其中,$g_i \in R^{1\times D}$ 是第 i 个关节点的特征表示,$\bar{g}_i \in R^{1\times H}$ 表示升维之后的第 i 个关节点的特征表示,H 为升维之后的维度,$W \in R^{D\times H}$ 表示线性投影层的权重,$\bar{G} \in R^{K\times H}$ 是 G 升维之后的特征表示。高维度的特征能够表示更丰富的信息,包括关键点之间的相对位置、关键点的运动趋势等,使模型能够更好地学习和理解人体姿态的复杂性和多样性。

然后,将 $\bar{G}$ 输入到 N 个 Mixer Layer 中以深度融合不同关节之间的特征,模型通过学习去组合不同的关节,最后由 Whole Split 模块拆分为 M 个子结构:

$$\bar{G}'=\mathrm{Mixer}(\bar{G}), \tag{4}$$

$$T=\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_M \end{bmatrix}=(\bar{G}'^{\mathrm{T}}W_1)^{\mathrm{T}}W_2, \tag{5}$$

其中,$\bar{G}' \in R^{K\times H}$ 表示经过 N 个 Mixer Layer 融合之后的特征,$W_1 \in R^{K\times M}$,$W_2 \in R^{H\times H}$ 表示全连接层的权重,$t_i \in R^{1\times H}$ 是第 i 个标记,H 表示标记的维度,t_i 近似对应于几个相互依赖的关节子结构,$T \in R^{M\times H}$ 表示拆分为子结构之后的整体姿态特征。t_i 聚

合了多个关节的特征，目的是突出每个部位的运动特征，高维度的特征包含了许多复杂的信息，如果将高维度的 t_i 输入到时序提取模块，x_t 的维度将为 $M\times H$，过高的维度将导致模型无法计算。因此，需要使用池化下采样进行降维，低维度的信息更能突出运动特征。

最后，将 M 个子结构特征输入到通道注意力机制模块——Squeeze - and - Excitation Networks (SENet)[18]中，以提高 M 个子结构中某些关键子结构的关注度，再使用平均池化将 M 个特征融合、降维，最后拼接起来得到整个骨架的特征：

$$S=\begin{bmatrix}s_1\\s_2\\\vdots\\s_M\end{bmatrix}=\text{Pooling}(\text{SENet}(T)),\tag{6}$$

$$x_t=\text{Concat}(S),\tag{7}$$

其中，$s_i\in R^{1\times 4}$ 是平均池化后的第 i 个标记特征；$S\in R^{M\times 4}$ 表示池化之后的整体姿态特征；$x_t\in R^{1\times 4M}$ 是最后的特征表示，融合了整个骨架的信息。

CTMM 的分解方式允许其在对人体动作进行建模时更精细地关注不同部位的特征。通过平均池化的下采样操作，得到了对人体多个部位状态的低维表示，从而能够捕捉人体运动的整体信息。下采样能够抹平坐标分布偏移带来的一些不规则性的影响，减少冗余信息，使特征表达更加稳定和鲁棒。

算法 1：CTMM 算法

输入：当前时刻的人体 K 个关节坐标矩阵 G

输出：当前时刻的骨架整体姿态特征 x_t

说明：f 为线性投影层

1：begin

2：　$\bar{G}\leftarrow f_1(G)$

3：　for $i\leftarrow 1$ to N do　//Mixer Layer 共有 N 层

4：　　$\bar{G}\leftarrow \text{Layer Norm}(\bar{G})$

5：　　$\bar{G}_1\leftarrow f_2^i(\bar{G}^T)$

6：　　$\bar{G}_2\leftarrow \bar{G}_1^T+\bar{G}$

7：　　$\bar{G}_2\leftarrow \text{Layer Norm}(\bar{G}_2)$

8：　　$\bar{G}_2\leftarrow f_3^i(\bar{G}_2)$

9：　　$\bar{G}\leftarrow \bar{G}+\bar{G}_1+\bar{G}_2$

10：　end

11：　$\bar{G}\leftarrow \text{Layer Norm}(\bar{G})$　//将特征进行归一化

12：　$T\leftarrow \text{Split}(\bar{G})$　//使用拆分模块将骨架拆分为 M 个子结构

13：　$S\leftarrow \text{Average Pooling}(\text{SENet}(T))$　//对 M 个子结构进行下采样

14：　$x_t\leftarrow \text{Concat}(S)$　//将 M 个子结构特征拼接得到整个骨架特征

15：end

2.2　时序特征提取模块——Attention LSTM

Attention LSTM 结构如图 5 所示。与 Mogrifier LSTM 类似，Attention LSTM 不改变 LSTM 的基本结构，而是通过新的门控机制，让 x_t 和 h_{t-1} 分别进行多层注意力计算，再送入 LSTM 单元。注意力模块层数由超参数 r 控制，在更新 x_t 与 h_{t-1} 时分别计算两者不同的 Q、K、V，经过注意力计算之后得到 O。此外，加入跳跃连接，将 O 与原始值 x_t 相加并进行归一化之后送入前馈神经网络得到新的 x_t 与 h_{t-1}。前馈神经网络可以使模型学习到更高级别的特征表示，从而提升模型的特征表达能力。Attention LSTM 中具体的计算过程如下。

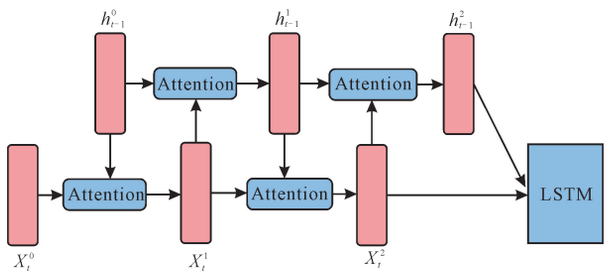

图 5　Attention LSTM 结构($r=4$)

Fig. 5　Structure of the Attention LSTM ($r=4$)

x_t 与 h_{t-1} 的更新是交替进行的，模型首先会更新 x_t，再更新 h_{t-1}，具体见图 6。

在更新将 x_t 时，将 x_t 作为 K，h_{t-1} 分别乘以两个不同的权重得到 Q 和 V，然后输入到注意力模块中进行计算。计算公式为

$$K=x_t^{i-1},i\in[1,\cdots,\lfloor(r+1)/2\rfloor],\tag{8}$$

$$Q=W_{q1}^i h_{t-1}^{i-1},i\in[1,\cdots,\lfloor(r+1)/2\rfloor],\tag{9}$$

$$V=W_{v1}^i h_{t-1}^{i-1},i\in[1,\cdots,\lfloor(r+1)/2\rfloor],\tag{10}$$

$$O=\text{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_K}}\right)V,i\in[1,\cdots,\lfloor(r+1)/2\rfloor],\tag{11}$$

$$x_t^i=\text{FFN}(\text{Layer Norm}(O+x_t^{i-1})),i\in[1,\cdots,\lfloor(r+1)/2\rfloor],\tag{12}$$

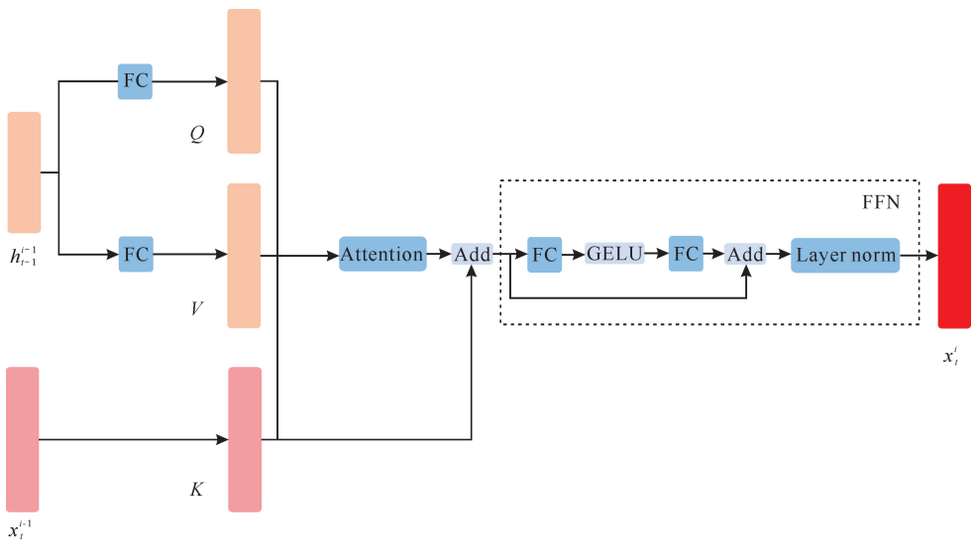

图 6　x 的更新过程

Fig. 6　Update process of x

其中，$r \in N$ 代表注意力层的层数，当 $r=0$ 时，为原始 LSTM；在更新 h_{t-1} 时，将 h_{t-1} 作为 K，x_t 分别乘以两个不同的权重得到 Q 和 V，计算过程同上，区别在于当 r 是一个奇数时，x_t 会比 h_{t-1} 的更新次数多一次，K、Q、V 与当前输入 x_t 与 h_{t-1} 相关。这种注意力机制的引入增强了模型对序列中重要信息的关注程度，有助于区分相似手势并最大程度地保留关键细节，从而在交警手势识别任务中表现出更好的性能。

算法 2：Attention LSTM 算法

输入：LSTM 的上一时刻输出的隐藏状态 h_{t-1}，细胞状态 c_{t-1} 以及 CTMM 提取当前时刻的特征 x_t

输出：当前时刻的手势类别 $f(h_t)$

说明：f_m 为线性投影层

1：begin

2：　for $i \leftarrow 1$ to r do　//注意力模块层数共有 N 层

3：　　if $i\%2!=0$ then　//更新 x_t

4：　　$K_i \leftarrow x_t$

5：　　$Q_i \leftarrow W_{q1}^i h_{t-1}$

6：　　$V_i \leftarrow W_{v1}^i h_{t-1}$

7：　　$O_i \leftarrow \mathrm{Softmax}(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_k}})V_i$　//进行注意力计算

8：　　$x_t \leftarrow \mathrm{FFN}(\mathrm{Layer\ Norm}(O_i + K_i))$　//将特征输入到前馈神经网络

9：　　else　//更新 h_{t-1}

10：　　$K_i \leftarrow h_{t-1}$

11：　　$Q_i \leftarrow W_{q2}^i x_t$

12：　　$V_i \leftarrow W_{v2}^i x_t$

13：　　$O_i \leftarrow \mathrm{Softmax}(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_k}})V_i$　//进行注意力计算

14：　　$h_{t-1} \leftarrow \mathrm{FFN}(\mathrm{Layer\ Norm}(O_i + K_i))$　//将特征输入到前馈神经网络

15：　end

16：　$(h_t, c_t) \leftarrow \mathrm{LSTM}(x_t, (h_{t-1}, c_{t-1}))$　//将新的 x_t 和 h_{t-1} 输入到 LSTM 单元

17：end

3　实验与结果分析

3.1　实验设置

3.1.1　中国交警手势数据集

公开的中国交警手势数据集包含 21 个视频样本，每个视频长度为 5－10 min，帧速率为 15 FPS，并涵盖了不同的场景。最初的数据集划分如下：训练集包含 11 个视频，共计 1 789 个手势；测试集包含 10 个视频，共计 1 565 个手势。虽然数据集提供了每个视频样本每一帧的真实标签序列，但由于存在一些标注错误，本研究对数据集进行了重新整理和修正。重新整理后的数据集划分为训练集 10 个视频，共计 1 645 个手势；测试集 11 个视频，共计 1 709 个手势。实验中使用 PCT 提取每一帧的人体骨架数据，并将这些数据保存为.pkl 文件以便后续训练和测试。实

验采用的深度学习框架为 PyTorch 2.1.0,GPU 为配备 12 GB 显存的 Nvidia GeForce GTX 1080Ti。优化器为 Adam,初始学习率为 0.000 1。训练过程中,epoch 设置为 6 000,batch size 设置为 10。在每次迭代时,从训练集的 10 个视频中随机截取连续的 90 帧,将每个截取帧序列的前 11 帧的标签屏蔽,并选择交叉熵作为反向传播的损失函数。测试阶段,batch size 设置为 1,将所有测试数据逐帧输入模型进行预测,符合实时推理的标准。

3.1.2 手语数据集

本研究引用阿根廷手语数据集 LSA64、WLASL100 和 CSL-500 3 个流行的开放单词级手语数据集进行实验。这些数据集的具体信息如表 1 所示。文献[19]提供了 LSA64 和 WLASL100 数据集的骨架数据,其中作者使用 Vision API 的标准姿态估计算法从每个视频帧中获取头部、身体和手部一共 54 个关节的二维坐标。为了统一处理数据集,实验中对 3 个数据集的每个视频取前面 70 帧,对于不足 70 帧的样本,在后面填充 0 使其满足 70 帧;对于超过 70 帧的样本,则截取前面 70 帧。实验使用的深度学习框架为 PyTorch 2.1.0,GPU 为配备 12 GB 显存的 Nvidia GeForce GTX 1080Ti。优化器选择 Adam,初始学习率为 0.000 1。batch size 设置为 50,epoch 设置为 300。预测每个视频的标签时,选择交叉熵作为反向传播的损失函数。LSA64 和 CSL-500 数据集按 8∶2 的比例随机划分为与表演者无关的训练集和测试集。各个手语数据集的信息如表 1 所示。

表 1 手语数据集信息

Table 1 Sign language dataset information

数据集 Dataset	训练集总数 Training sets	测试集总数 Test sets	手势类别数 Number of classes	关节数 Number of joints
LAS64	2 560	640	64	54
WLASL100	1 442	258	100	54
CSL-500	10 000	2 500	500	25

3.1.3 评价指标

按照训练的方式不同,本研究采用两种不同的评价指标来评估算法性能,一种是编辑距离(Edit-distance)[12],一种是普通的准确率计算方式。手语数据集采用的评价指标为普通的准确率计算方式。编辑距离是一种用于衡量两个字符串之间差异程度的量化方法。它衡量的是将一个序列转变成另一个序列所需的最少编辑操作次数。这些编辑操作包括替换、插入、删除。式(13)为编辑距离的计算公式,式(14)为普通准确率的计算公式。

$$\text{Distance}=\frac{(N-I-D-S)}{N}, \tag{13}$$

$$\text{Accuracy}=\frac{P}{N}, \tag{14}$$

其中,N 为视频中真实手势的总数,I 是视频中插入手势的总数,D 是系统中删除手势的总数,S 是系统中替换手势的总数,P 是预测正确的手势数。使用训练好的模型分别对测试集的每个视频的每一帧进行预测,然后将预测序列与标注序列进行编辑距离准确率计算。编辑距离准确率越高,表示模型预测的序列与真实序列之间的差异越小,模型的性能越好。每个视频代表不同的场景(室内、室外、光线不足、背景复杂等),实验结果汇总于表 2 中,用于评估模型在不同场景下的性能表现。

3.1.4 PoseLSTM 在不同任务中的超参数设置

PoseLSTM 在不同的任务中设置了不同的超参数,如表 2 所示。其中,K 为关节点数量,M 为 CTMM 提取的子结构数量,R 为 Attention LSTM 中注意力模块的层数,N 为 CTMM 中 Mixer Layer 的层数,Pooling Size 为自适应池化层的参数,LSTM Input 为 LSTM 输入的特征维度,LSTM Hidden 为隐藏层的维度,CTPG 表示中国交警手势数据集。

表 2 不同任务中模型的超参数设置及大小

Table 2 Hyperparameter settings and sizes for models in different tasks

数据集 Dataset	K	M	R	N	池化大小 Pooling size	LSTM 输入特征维度 Dimension of LSTM input features	LSTM 隐藏层维度 Dimension of LSTM hidden layers	参数量/M Params/M
CTPG	17	64	2	4	4	256	64	5.30
LSA64	54	64	4	4	4	256	128	9.70
WLASL100	54	128	4	4	4	512	128	23.90
CLS-500	25	64	7	4	4	256	128	14.60

由表 2 可知,PoseLSTM 在 4 个数据集上的参数量都不是很大,模型针对不同任务的超参数可以根据实际情况进行人工调整。一般来说,提取的身体关节点越少,模型的超参数就越小,模型也就越小。中国交警通过身体与手臂的摆动传递交通信号,因此交警手势的识别不需要关注手掌的特征,提取的关节点较少;而手语的表达需要关注手指的运动状态,因此手语数据集提取的关节点较多。WLASL100 数据集相对于其他数据集更为复杂,并且数据量较小,较小的参数量无法捕捉到更细致的特征,导致小模型准确率较低,因此增加模型参数量。对于其他任务,模型的 M 参数固定为 64,以确保模型的轻量化。

3.2 结果分析

3.2.1 在中国交警手势数据集上的对比实验

PoseLSTM 在各个测试视频中的编辑距离如表 3 所示,编辑距离均为 1.0。这表明 PoseLSTM 即使在多种不同的场景下(如室内、室外、光线不足、背景复杂以及存在车辆运动等)仍然能够保持非常高的准确率。此外,训练完成后的模型参数量仅为 5.3 M,表明模型具有非常强的特征捕捉能力和鲁棒性。

表 3 在各个测试视频的编辑距离

Table 3 Edit-distance in various test videos

文件 File	预测数 Prediction	N	S	D	I	编辑距离 Edit-distance
104.mp4	161	161	0	0	0	1.0
012.mp4	177	177	0	0	0	1.0
101.mp4	161	161	0	0	0	1.0
010.mp4	145	145	0	0	0	1.0
008.mp4	161	161	0	0	0	1.0
016.mp4	161	161	0	0	0	1.0
002.mp4	161	161	0	0	0	1.0
004.mp4	99	99	0	0	0	1.0
018.mp4	161	161	0	0	0	1.0
102.mp4	161	161	0	0	0	1.0
014.mp4	161	161	0	0	0	1.0
Total	1 709	1 709	0	0	0	1.0

PoseLSTM 与其他几种相关模型的比较如表 4 所示。基于 2D 人体姿态的交警手势识别方法的准确率普遍较高,这是因为基于 RGB 的方法容易受到室外复杂背景的影响,而基于人体姿势的方法能够较好地避免这种影响。PoseLSTM 在中国交警手势数据集的两种评价指标中都获得了 100%的准确率,超过了目前的基线方法。PoseLSTM 的高性能得益于其独特的架构设计,从而使 PoseLSTM 的性能超过了基于 GCN 的模型。采用编辑距离的模型实验结果均来自文献[11],其余模型的结果均来自文献[8]。

表 4 在中国交警手势数据集上的对比实验

Table 4 Comparative experiment on gesture dataset of Chinese traffic police

模型 Models	基于 RGB RGB-based	基于姿态 Pose-based	编辑距离 Edit-distance	准确率/% Accuracy/%
ConvLSTM[20]	√	×		82.40
VGGNet-SSD+KEN+LSTM[21]	√	×	87.04	
C3D-Net[22]	√	×		87.12
P3D ResNet[23]	√	×		92.21
KEN+LSTM[12]	×	√	91.18	

续表

Continued table

模型 Models	基于 RGB RGB-based	基于姿态 Pose-based	编辑距离 Edit distance	准确率/% Accuracy/%
KEN+Bidirectional LSTM[24]	×	√	91.04	
ResNet Part Localizer+LSTM[25]	×	√	87.22	
DenseNet Part Localizer+LSTM[26]	×	√	89.66	
Convolutional LSTM[27]	×	√	80.77	
GRSCTFF[11]	×	√	94.12	
MD-GCN[8]	×	√		98.95
PoseLSTM (this study)	×	√	**100.00**	**100.00**

Note:× indicates that the corresponding modality is not used,and √ indicates that the corresponding modality is used;bold indicates the best result.

3.2.2 在 LSA64 数据集上的对比实验

表 5 为 PoseLSTM 在 LSA64 数据集上与其他相关模型的对比结果。PoseLSTM 取得了 100.00% 的准确率,实现了与当前最优算法相同的性能,SPOTER[19]使用 Transformer 和独特数据增强算法增强了其鲁棒性,在 LSA64 数据集上的准确率达到了 100.00%。

表 5 在 LSA64 数据集上的对比实验

Table 5 Comparison experiments on the LSA64 dataset

模型 Models	基于 RGB RGB-based	基于姿态 Pose-based	准确率/% Accuracy/%
ELM+MN CNN[28]	√	√	97.81
LSTM+DSC[29]	√	√	99.84
SwC GRMMixer[30]	√	×	98.75
R(2+1)D[31]	√	×	99.53
MEMP[32]	√	×	99.06
LSTM+LDS[33]	√	×	98.09
I3D[34]	√	×	98.91
SPOTER[19]	×	√	100.00
PoseLSTM (this study)	×	√	100.00

Note:× indicates that the corresponding modality is not used,and √ indicates that the corresponding modality is used.

3.2.3 在 WLASL100 数据集上的对比实验

PoseLSTM 在 LSA64 数据集上与其他相关模型的对比结果如表 6 所示。基于 RGB 的模型优于基于姿态的模型,这可能是因为 WLASL100 数据集中手语关节点的遮挡问题导致姿态估计不准确,从而使交警手势识别模型难以捕捉到细节特征,影响了对手势的区分能力。PoseLSTM 的准确率为 59.69%,这一较低的准确率主要是由于 PoseLSTM 需要大量数据来学习更为泛化的特征和模式。然而,WLASL100 数据集的样本量较少,限制了模型的特征学习能力。相比之下,采用数据增强算法的 Transformer 模型在识别准确率上表现更好。

表 6 在 WLASL100 数据集上的对比实验

Table 6 Comparison experiments on the WLASL100 dataset

模型 Models	基于 RGB RGB-based	基于姿态 Pose-based	准确率/% Accuracy/%
I3D[34]	√	×	65.89
TK-3D ConvNet[35]	√	×	77.55
Fusion-3[36]	√	×	75.67
GCN-BERT[37]	×	√	60.15
Pose-TGCN[34]	×	√	55.43
Pose-GRU[34]	×	√	46.51
SPOTER[19]	×	√	63.18
PoseLSTM (this study)	×	√	59.69

Note:× indicates that the corresponding modality is not used,and √ indicates that the corresponding modality is used.

3.2.4 在 CLS-500 数据集上的对比实验

表 7 为 PoseLSTM 在 CLS-500 数据集上与其他相关模型的对比结果。即使是在具有 500 个类别的大型手语数据集上,PoseLSTM 也取得了较好的性能,证明了 PoseLSTM 的有效性。

表 7 在 CLS-500 数据集上的对比实验

Table 7 Comparison experiments on the CLS-500 dataset

模型 Models	基于 RGB RGB-based	基于姿态 Pose-based	准确率/% Accuracy/%
R(2+1)D[38]	√	×	97.45
3D CNN[39]	√	×	88.70
HTAN[40]	√	×	93.10
Res-C3D[41]	√	×	89.20
SwC GRMMixer[30]	√	×	98.54
PoseLSTM (this study)	×	√	96.40

Note:× indicates that the corresponding modality is not used,and √ indicates that the corresponding modality is used.

3.2.5　鲁棒性实验

为了测试模型的鲁棒性,可以通过减少输入中的一部分关节点来观察这种扰动将如何影响最终精度。通过在训练和测试过程中每帧随机丢弃一个身体关节点(丢弃概率 p)引入扰动,实验在中国交警手势数据集上进行。如表 8 所示,PoseLSTM 在 $p=1$ 的情况下准确率仅下降 1.69%,这表明模型具有较强的鲁棒性。

表 8　不同丢弃概率的识别性能

Table 8　Recognition performance of different dropout probabilities

丢弃概率 p	准确率/% Accuracy/%
0	100.00
1/8	100.00
1/4	99.60
1/2	98.80
1	98.31

3.2.6　消融实验

为了验证 Attention LSTM 中注意力门控的有效性,将 PoseLSTM 与剔除注意力门控的 LSTM 以及 Mogrifier LSTM 进行对比实验,实验在中国交警手势数据集和 3 个手语数据集上进行。为了确保公平性和一致性,所有模型的超参数全部保持一致。如表 9 所示,在中国交警手势数据集上,去掉注意力门控后,LSTM 的准确率只有 97.22%,经过改进的 Mogrifier LSTM 的准确率为 98.44%。在手语数据集中,Mogrifier LSTM 的性能甚至不如普通的 LSTM。这是因为手语数据集相似手势比较多,Mogrifier LSTM 的计算方式破坏了 LSTM 中输入和隐藏状态的原始信息,未能有效地区分这些相似手势,导致性能下降。Attention LSTM 在 4 个数据集上的性能显著优于 LSTM 和 Mogrifier LSTM,表明其注意力门控的有效性,其强大的信息传递能力在长视频序列任务中展现出更好的性能。

为验证 PoseLSTM 中不同参数设置对模型性能的影响,对 Attention LSTM 的 x_t、h_{t-1} 注意力模块层数 r 和 CTMM 的 Mixer Layer 层数 N 取不同值,测试不同值对模型性能的影响。为了确保实验的公平性,当改变 r 值时,Mixer Layer 的层数 N 固定为 4;当改变 N 的值时,r 值固定为 4。将 WLASL100 与 CSL-500 两个手语数据集作为实验数据集,图 7 展示了当 $r\in\{2,3,\cdots,8\}$ 时,两个数据集测试集的准确率;图 8 展示了 N 取值为 2－8 时,两个数据集测试集的准确率。

表 9　在 4 个数据集上的消融实验结果

Table 9　Results of ablation experiments on four datasets

数据集 Dataset	模型 Method	准确率/% Accuracy/%
CTPG	CTMM+LSTM	97.22
	CTMM+Mogrifier LSTM	98.44
	CTMM+Attention LSTM	100.00
LSA64	CTMM+LSTM	99.84
	CTMM+Mogrifier LSTM	99.37
	CTMM+Attention LSTM	100.00
WLASL100	CTMM+LSTM	35.20
	CTMM+Mogrifier LSTM	32.94
	CTMM+Attention LSTM	59.69
CSL-500	CTMM+LSTM	87.33
	CTMM+Mogrifier LSTM	86.00
	CTMM+Attention LSTM	96.40

Note:CTPG is the Chinese Traffic Police Gesture dataset.

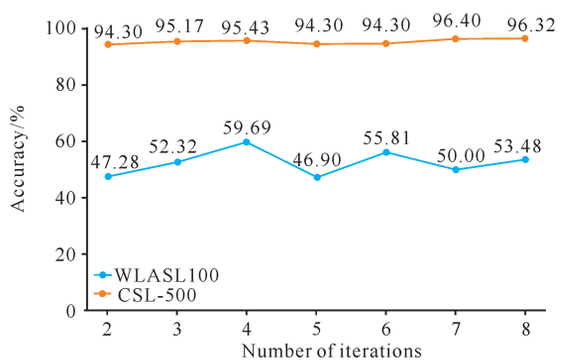

图 7　迭代次数 r 对实验结果的影响

Fig. 7　Influence of iteration times r on experimental results

如图 7 所示,在 WLASL100 数据集中,当 $r=4$ 时取得最佳的结果,而在 CSL-500 数据集中,在 $r=7$ 时取得最好的结果。这表明对于不同的任务,要取得最好的性能,r 的取值需要变化。值得注意的是,在 $r=4$ 时,两个数据集都取得了较为优异的结果,持续增大 r 值可能会带来性能的进一步提升,但同时也会显著增加计算成本。特别是在 CSL-500 数据集上,不同 r 之间的性能差异小于 1%,因此单纯为了较小的性能增益而增加 r 值可能是不经济的。

如图 8 所示,在 CSL-500 数据集 N 的不同取值中,模型性能的波动变化不大,在 $N=8$ 时达到最大

值,其次是 $N=4$;而对于 WLASL100 数据集,在 $N=4$ 时,模型性能最好。因此,考虑到计算成本和误差的影响,在 $N=4$ 时为最优平衡点。

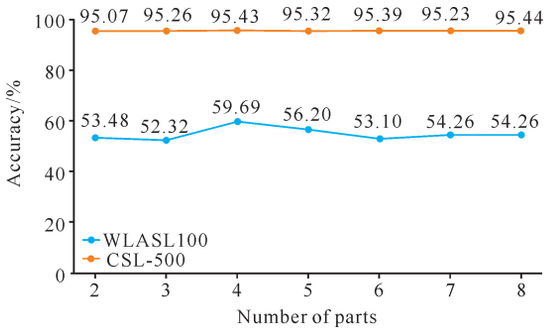

图 8　子结构数量 N 对实验效果的影响

Fig. 8　Influence of the number of substructures N on experimental results

4　结论

本研究提出一种基于 2D 人体姿态的轻量级交警手势识别模型 PoseLSTM,提高了中国交警手势数据集的识别准确率。PoseLSTM 中的骨架特征组合编码器 CTMM 有效地提取了人体骨架特征,且无需任何数据增强技术,解决了现有算法无法充分捕捉骨架关节坐标间运动特征的问题。CTMM 提取到的特征被输入到 Attention LSTM 模块中,Attention LSTM 能够建立视频序列中更强的长距离依赖关系,解决了 LSTM 潜在的性能瓶颈问题,提升了 LSTM 在视频序列任务中的性能。与当前普遍采用的基于 GCN 的模型相比,PoseLSTM 展现了更优的视频序列信息处理能力。在实时动态交警手势识别中,基于 GCN 的模型通常需要一组完整的视频帧才能识别一个手势,而 PoseLSTM 可以实现单帧输入,并与之前的帧进行信息交互以识别每一帧的手势,显著降低了计算消耗。PoseLSTM 在 3 个单词级别的手语数据集上的实验也实现了较好的性能。无论是在大型数据集还是小型数据集上,PoseLSTM 都表现出较好的泛化能力,并在中国交警手势数据集中取得了最先进的性能。这些结果证明了 PoseLSTM 的有效性和实用性。在未来的工作中,将继续研究如何将 PoseLSTM 应用于视频动作识别领域,以验证其在不同条件下的表现,提高模型的实用性。

参考文献

[1]　中共中央,国务院.国家综合立体交通网规划纲要[EB/OL].(2021-02-24)[2024-06-02].https://www.gov.cn/zhengce/2021-02/24/content_5588654.htm.

[2]　刘永涛,刘永杰,孙斐然,等.基于深度学习的交警动态手势检测与识别方法研究[J].武汉理工大学学报(交通科学与工程版),2024,48(3):441-447.

[3]　程贝芝,伍鹏,寇静雯,等.结合全局上下文信息的交警手势识别方法[J].中南民族大学学报(自然科学版),2023,42(3):349-356.

[4]　SHI P C,ZHANG Q,YANG A X. Dual-module spatial temporal information enhancement graph convolutional network for recognizing traffic police command gestures [J]. Signal,Image and Video Processing,2025,19:92.

[5]　FU Z,CHEN J J,JIANG K,et al. Traffic police 3D gesture recognition based on spatial-temporal fully adaptive graph convolutional network [J]. IEEE Transactions on Intelligent Transportation Systems,2023,24(9):9518-9531.

[6]　KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks [C]//5th International Conference on Learning Representations (ICLR). Toulon:ICLR,2017:2713-2727.

[7]　YAN S J,XIONG Y J,LIN D H,et al. Spatial temporal graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New York:ACM,2018:7444-7452.

[8]　XIONG X,WU H Y,MIN W D,et al. Traffic police gesture recognition based on gesture skeleton extractor and multichannel dilated graph convolution network [J]. Electronics,2021,10(5):551.

[9]　LIU K,ZHENG Y,YANG J Y,et al. Chinese traffic police gesture recognition based on graph convolutional network in natural scene [J]. Applied Sciences,2021,11(24):11951.

[10]　ZHU D Y,ZHANG Z W,CUI P,et al. Robust graph convolutional networks against adversarial attacks [C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York:ACM,2019:1399-1407.

[11]　何坚,廖俊杰,张丞,等.基于长短时记忆和深度神经网络的视觉手势识别技术[J].图学学报,2020,41(3):372-381.

[12]　HE J,ZHANG C,HE X L,et al. Visual recognition of traffic police gestures with convolutional pose machine and handcrafted features [J]. Neurocomputing,2020,390:248-259.

[13] MELIS G,KOCISKY T,BLUNSOM P. Mogrifier LSTM [C]//8th International Conference on Learning Representations (ICLR). Virtual: ICLR, 2020: 13316-13329.

[14] GENG Z G,WANG C Y,WEI Y X,et al. Human pose as compositional tokens [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 660-671.

[15] LIU Z Y,ZHANG H W,CHEN Z H,et al. Disentangling and unifying graph convolutions for skeleton-based action recognition [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2020: 140-149.

[16] LI M S,CHEN S H,CHEN X,et al. Actional-structural graph convolutional networks for skeleton-based action recognition [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE,2019:3595-3603.

[17] NIE B X H,XIONG C M,ZHU S C. Joint action recognition and pose estimation from video [C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015:1293-1301.

[18] HU J,SHEN L,SUN G. Squeeze-and-excitation networks [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York:IEEE,2018:7132-7141.

[19] BOHACEK M,HRUZ M. Sign pose-based transformer for word-level sign language recognition [C]//Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). New York:IEEE,2022:182-191.

[20] SHI X J,CHEN Z R,WANG H,et al. Convolutional LSTM network:a machine learning approach for precipitation nowcasting [C]//Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge:MIT Press,2015:802-810.

[21] LIU W,ANGUELOV D,ERHAN D,et al. SSD:single shot MultiBox detector [M]//Proceedings of the Computer Vision-ECCV 2016. Cham:Springer International Publishing,2016:21-37.

[22] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). New York:IEEE,2015:4489-4497.

[23] QIU Z F,YAO T,MEI T. Learning spatio-temporal representation with pseudo-3D residual networks [C]//Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). IEEE,2017: 5534-5542.

[24] PIGOU L,VAN DEN OORD A,DIELEMAN S,et al. Beyond temporal pooling:recurrence and temporal convolutions for gesture recognition in video [J]. International Journal of Computer Vision,2018,126(2):430-439.

[25] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition [C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2016: 770-778.

[26] HUANG G,LIU Z,VAN DER MAATEN L,et al. Densely connected convolutional networks [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE,2017:2261-2269.

[27] JI S W,XU W,YANG M,et al. 3D convolutional neural networks for human action recognition [C]//Proceedings of the 27th International Conference on International Conference on Machine Learning. New York: ACM,2010:495-502.

[28] IMRAN J,RAMAN B. Deep motion templates and extreme learning machine for sign language recognition [J]. The Visual Computer,2020,36(6):1233-1246.

[29] KONSTANTINIDIS D,DIMITROPOULOS K,DARAS P. A deep learning approach for analyzing video and skeletal features in sign language recognition [C]//Proceedings of the 2018 IEEE International Conference on Imaging Systems and Techniques (IST). New York:IEEE,2018:1-6.

[30] LIU T Y,TAO T F,ZHAO Y Z,et al. A signer-independent sign language recognition method for the single-frequency dataset [J]. Neurocomputing,2024,582: 127479.

[31] MARAIS M,BROWN D,CONNAN J,et al. Spatio-temporal convolutions and video vision transformers for signer-independent sign language recognition [C]//Proceedings of the 2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). New York:IEEE,2023:1-6.

[32] ZHANG X Y,LI X Q. Dynamic gesture recognition

based on MEMP network [J]. Future Internet, 2019, 11(4):91.

[33] KONSTANTINIDIS D, DIMITROPOULOS K, DARAS P. Sign language recognition based on hand and body skeletal data [C]//Proceedings of the 2018 - 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON). New York: IEEE, 2018:1-4.

[34] LI D X, OPAZO C R, YU X, et al. Word-level deep sign language recognition from video: a new large-scale dataset and methods comparison [C]//Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2020: 1448-1458.

[35] LI D X, YU X, XU C C, et al. Transferring cross-domain knowledge for video sign language recognition [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: New York: 6205-6214.

[36] HOSAIN A A, SELVAM SANTHALINGAM P, PATHAK P, et al. Hand pose guided 3D pooling for word-level sign language recognition [C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). New York: IEEE, 2021: 3428-3438.

[37] TUNGA A, NUTHALAPATI S V, WACHS J. Pose-based sign language recognition using GCN and BERT [C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW). New York: IEEE, 2021:31-40.

[38] HAN X Z, LU F, YIN J Q, et al. Sign language recognition based on R(2+1)D with spatial - temporal - channel attention [J]. IEEE Transactions on Human-Machine Systems, 2022, 52(4):687-698.

[39] HUANG J, ZHOU W G, LI H Q, et al. Attention-based 3D-CNNs for large-vocabulary sign language recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(9):2822-2832.

[40] 黄杰. 基于深度学习的手语识别技术研究[D]. 合肥:中国科学技术大学, 2018.

[41] ZHANG S J, ZHANG Q. Sign language recognition based on global-local attention[J]. Journal of Visual Communication and Image Representation, 2021, 80: 103280.

A Novel Chinese Traffic Police Gesture Recognition Network Based on Human Pose

QIN Xiao[1,2], LI Yongyu[1], WU Kunsheng[3], YUAN Chang'an[4**], TAN Sijing[1], LIU Shanrui[1]

(1. Guangxi Key Laboratory of Human-Computer Interaction and Intelligent Decision Making, Nanning Normal University, Nanning, Guangxi, 530100, China; 2. Guangxi Regional Collaborative Innovation Center for Multi-Source Data Integration and Intelligent Processing, Guilin, Guangxi, 541004, China; 3. Nanning Arboretum, Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530225, China; 4. Guangxi Academy of Sciences, Nanning, Guangxi, 530012, China)

Abstract: Traffic police gesture recognition is crucial for autonomous driving technology. Existing traffic police gesture recognition methods based on human pose have shortcomings such as incomplete feature extraction and insufficient robustness in skeleton feature extraction. The extraction of time series features suffers from dynamic information loss, weak time series dependency, and poor real-time performance. Additionally, the extraction effectiveness is easily affected by environmental backgrounds. In view of the shortcomings above, a novel traffic police gesture recognition network Pose Long Short-Term Memory (PoseLSTM) based on human pose is proposed. Specifically, the Compositional Tokens Multi-layer Perceptron Mixer (CTMM) in PoseLSTM can capture the relational features between body joints and transform the joint information through dependency modeling to form multi-part feature representations, addressing the problem of Long

Shore-Term Memory (LSTM)-based algorithms failing to effectively extract skeleton features. Moreover, the hybrid architecture Attention LSTM in PoseLSTM can better integrate input and hidden state information, outperforming the original LSTM. Experimental results showed that PoseLSTM achieved the accuracy of 100.00% on the open-source Chinese traffic police gesture dataset, achieving optimal performance. Furthermore, to demonstrate the generalization ability of PoseLSTM, experiments were conducted on open sign language datasets LSA64, WLASL-100, and CSL-500, with the accuracy of 100.00%, 59.69%, and 96.40%, respectively.

Key words: traffic police gesture recognition; attention mechanism; LSTM; joint combination

责任编辑:陆　雁