空间上下文与时序特征融合的交警指挥手势识别技术

张 丞,何 坚,王伟东
(北京工业大学信息学部,北京 100124)

摘 要: 针对无人驾驶汽车快速准确识别交警指挥手势的需求,本文在分析交警指挥手势的关节铰接特征基础上,建立基于关节点和骨架的交警指挥手势模型;其次,引入卷积姿势机(Convolutional Pose Machine,CPM)提取交警指挥手势的关键节点,进而提取交警指挥手势中骨架的相对长度及其与重力加速度的夹角作为空间上下文特征,并引入长短时记忆网络(Long Short Term Memory,LSTM)提取交警指挥手势的时序特征;最后,设计了融合空间上下文和时序特征的交警指挥手势识别机(Chinese Traffic Police Gesture Recognizer,CTPGR),创建了包含8种交警指挥手势、时长约2小时的交警指挥手势视频库对CTPGR进行训练验证,并通过实验将CTPGR与已有交警手势识别算法进行了对比分析. 实验证明CTPGR可以快速准确地识别交警指挥手势,系统对复杂背景和动态交警指挥手势具有较强的适应能力.

关键词: 交警指挥手势; 手势识别; 卷积姿势机; 长短时记忆; 特征提取
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112 (2020)05-0966-09
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.05.018

Visual Recognition of Chinese Traffic Police Gestures Based on Spatial Context and Temporal Features

ZHANG Cheng, HE Jian, WANG Wei-dong
(*Faculty of Information, Beijing University of Technology, Beijing* 100124, *China*)

Abstract: According to the need for driver assistance systems and intelligent vehicles to quickly and accurately identify traffic police command gestures, the articulated features of traffic police gesture is firstly analyzed, and a model based on the key points and skeletons of the police gesture is established. Secondly, the convolutional posture machine (CPM) is introduced to extract the key points of the traffic police gesture. Then the relative lengths of the gesture skeletons and the angles between each skeleton w. r. t. gravity are extracted as the spatial context features of the traffic police gesture. Meanwhile, long-term memory (LSTM) is introduced to extract the temporal features of traffic police gestures. Finally, the Chinese traffic police gesture recognizer (CTPGR) based on CPM and LSTM is designed, and a two-hour traffic police gesture video is recorded to train and verify the CTPGR. Experimental results show that the CTPGR is capable of recognizing traffic police gestures with high accuracy, and is fast enough for online gesture prediction.

Key words: traffic police gestures; gesture recognition; convolutional pose machine; long short term memory; feature extraction

1 引言

近年来随着无人驾驶技术的发展,越来越多城市规定自动驾驶车辆必须能够识别交警手势. 国内外研究人员对交警手势识别的研究总体上可分为两类:基于可穿戴传感器的方法和基于机器视觉的方法. 基于可穿戴传感器的方法使用惯性传感器来测量交警手臂运动. 例如,Wang等[1]通过交警手臂上佩戴的两个加速度计来感知手臂运动数据,并据此分类识别交警手势. Yuan等[2]使用数字手套提取交警手势的运动数据,并

收稿日期:2019-07-10;修回日期: 2019-10-25;责任编辑:孙瑶
基金项目:国家自然科学基金(No. 61602016);北京市科技计划项目(No. D171100004017003)

进行分类识别. 基于可穿戴传感器的方法可获得较好的识别率,但此类方法需要交警佩戴额外的动作感知装置,增加了交警的使用负担;同时较高的成本也限制了这一方法的推广应用.

基于计算机视觉的手势识别方法具有使用方便、成本较低等优点,是近年来交警手势识别的研究热点. 杜友田等[3]发表了一篇基于视觉的人体运动识别综述,完整地描述了人体运动的类别,介绍了一些基于概率网络和文法的识别方法. Guo 等[4,5]使用最大覆盖方案来监测交警身体区域,并结合关节旋转角度和 Gabor 特征,应用二维主成分分析方法分类识别手势. 此方法当交警手臂与图像平面垂直时难以正确识别手势. Eichner 等[6,7]提出了一种评估静止图像中人体空间布局(即头部、躯干和手臂的位置)的方法,一旦检测到人体可通过对其位置和外观约束来定位上肢和手,并进一步分类识别交警手势. 此外,管业鹏[8]提出了基于时/空运动特征的复杂人机交互场景下的指势用户对象识别新方法. Kang 等[9]融合手势点位检测识别技术,面向视频游戏 Quake II 设计了 10 种上肢手势作为交互接口. 张友梅等[10]提出了一种根据 3D 关键点识别人体动作的方法,使用了开源的深度数据集. Kinect 可提取图像深度信息,并定位人体骨骼,因此其在人体姿态识别中得到广泛应用[11]. 例如,Guo 等融合静态和动态描述符,通过 Kinect 提取 8 种交警手势特征,并采用平均结构相似性指数来识别交警手势[12]. Le 等利用 Kinect 捕获交警手势的深度图像,并采用支持向量机分类识别交警手势[13]. Zhou 等采集手势的彩色深度图像,并应用超限学习机计算手势的梯度特征直方图来分类识别交警手势[14]. 虽然 Kinect 可提供较精确的人体部位,并提高手势识别的准确率,但其集成的深度传感器工作范围受限,当其与交警距离较远时手势识别准确率会大幅降低.

近年来随着深度学习在图像识别、自然语言处理等方面取得成功,研究人员探索将深度学习应用于手势识别和动作识别中[15]. 比如,Fragkiadaki 等构造 3 层长短时记忆网络(Long Short Term Memory,LSTM)进行人体活动检测识别[16]. 上述研究成果表明卷积网络可以有效提取人体活动的空间特征,而 LSTM 等循环卷积网络可以有效提取人体活动的时序关系. 受卷积姿势机(Convolutional Pose Machine,CPM)[17]和部件亲和字段(Part Affinity Fields,PAF)[18]及 LSTM 启发,本文针对基于视觉的交警手势识别易受背景和手势运行变化等影响的问题,构造交警手势识别机(Chinese Traffic Police Gesture Recognizer,CTPGR)提取交警手势的时空特征,实现交警手势的快速准确识别.

2　交警手势建模

由中国公安部制定的“新版交通手势信号”自 2007 年起开始使用,本套手势信号由交警单人完成,用以指挥、疏导交通,规范驾驶人员的交通行为.

2.1　交警手势特征分析

“新版交通手势信号”由图 1 展示的 8 种手势组成. 本文额外引入了“待机(Stand in Attention)”,用于表示交警不做手势时的状态[2]. 概括起来,交警手势是由交警的头、颈、四肢和手构成的关节铰链式姿态. 铰链式姿态系统包含了每个人体部件的位置信息. 本文参照 CPM 将人体关键点称作“部件”,但在引用其他文献或数据集时,保留相应文献中的名称. 因此,本文中“关节”、“人体关键节点”、“部件”均指代铰链式姿势结构所包含关键节点的位置.

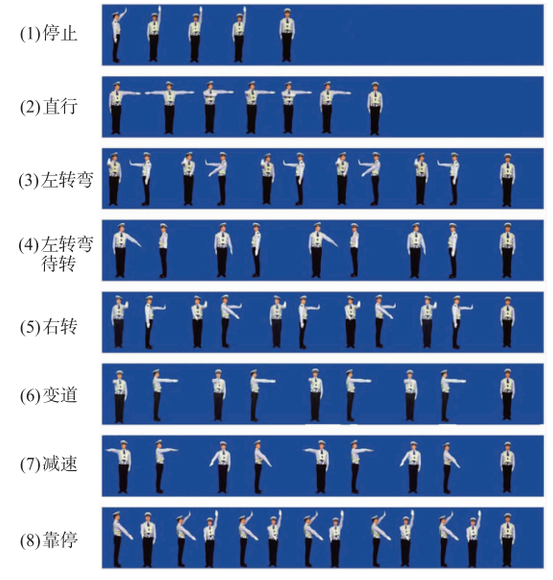


图1　交通手势信号动作

参考文献[19],交警的铰链式姿势可抽象为 14 个部件,如图 2(*a*)所示. 图 2(*b*)描述了这些部件的坐标,其集合为 $\boldsymbol{Y}$. $\boldsymbol{Y}$ 由头部关键节点 $\boldsymbol{Y}_{\text{head}}$、上身关键节点 $\boldsymbol{Y}_{\text{upper}}$、下身关键节点 $\boldsymbol{Y}_{\text{lower}}$三个集合构成,即

$$\boldsymbol{Y}=(Y_1,\cdots,Y_{14}),\boldsymbol{Y}_{\text{head}}=(Y_{13},Y_{14})$$
$$\boldsymbol{Y}_{\text{upper}}=(Y_1,\cdots,Y_6)\qquad \boldsymbol{Y}_{\text{lower}}=(Y_7,\cdots,Y_{12}) \tag{1}$$

依据解剖学中人体骨骼及相互间的依赖关系,$\boldsymbol{Y}$ 中相邻关键节点间存在连接依赖关系,这些连接依赖关系如图 2(*c*)所示. 交警手势所含关键节点间的连接关系集合表示为 $\boldsymbol{S}$. s 为其中的一条关键节点连接($s\in\boldsymbol{S}$),其起始关键节点和终止关键节点分别为 Y_m 和 Y_n,则 $\overrightarrow{Y_mY_n}$表示了交警手势所含的一条骨架矢量. 与关键节点分类方法类似,$\boldsymbol{S}$由头部骨架$\boldsymbol{S}_{\text{head}}$、上身骨架$\boldsymbol{S}_{\text{upper}}$和

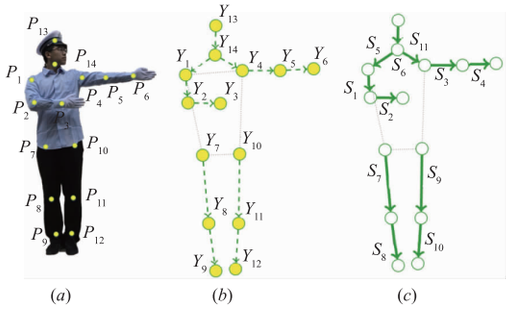


图2 基于关节点的交警手势模型

下身骨架 $\boldsymbol{S}_{\text{lower}}$ 三部分构成. 即

$$\begin{aligned}\boldsymbol{S}_{\text{head}} &= \{\overrightarrow{Y_{13}Y_{14}}\}\\ \boldsymbol{S}_{\text{upper}} &= \{\overrightarrow{Y_{14}Y_1},\overrightarrow{Y_1Y_2},\overrightarrow{Y_2Y_3},\overrightarrow{Y_{14}Y_4},\overrightarrow{Y_4Y_5},\overrightarrow{Y_5Y_6}\}\\ \boldsymbol{S}_{\text{lower}} &= \{\overrightarrow{Y_7Y_8},\overrightarrow{Y_8Y_9},\overrightarrow{Y_{10}Y_{11}},\overrightarrow{Y_{11}Y_{12}}\}\end{aligned} \tag{2}$$

为识别关键节点位置,本文引入并扩展 CPM,建立交警手势关键节点提取网络(Police Key-point Extracting Network,PKEN).

设 $\boldsymbol{Z}$ 为包含交警手势的图像上所有位置坐标(u,v)的集合,$\boldsymbol{Z}\subset\mathbf{R}^2$. 在图像中交警手势每个部件的位置用 Y_k 表示,$Y_k\in\boldsymbol{Z}$. 交警手势总共包含 14 个部件,因此 $\boldsymbol{Y}_k\in\{Y_1,\cdots,Y_{14}\}$. PKEN 由一系列多类预测器 g_t 组成,它们被训练用来预测同一图像在不同感受野下每个部件的位置. 具体而言,g_t 是一个分类器,下标 t 表示分类的阶段,每个阶段的感受野不同. g_t 预测该感受野下图像中点 z 属于部件 Y_k 的置信度,用 $b(Y_k=z)$ 表示置信度值. 这些 g_t 具有相同训练目标. 当 $t>1$ 时,g_t 是从图像位置 z 提取的特征值 x_z 和每个关键节点 Y_k 在 $t-1$ 时刻置信度的预测值的拼接函数. 即

$$g_t(x_z,b_{t-1})\rightarrow\{b_t^k(Y_k=z)\}_{k\in\{1,\cdots,14\}} \tag{3}$$

其中,x_z 为提取器 $\psi(\cdot)$ 在位置 z 提取的图像特征值. 即

$$\psi(z)\rightarrow x_z,z\in\boldsymbol{Z} \tag{4}$$

在分类器的第一阶段(即 $t=1$ 时),使用 x_z 表示图像位置 z 上的特征值,则分类器产生的值如下式(5)

$$g_1(X_z)\rightarrow\{b_1^k(Y_k=z)\}_{k\in\{1,\cdots,14\}} \tag{5}$$

其中,$b_1^k(Y_k=z)$ 表示图像中坐标点 z 属于部件 k 的置信度. 在 $t(t>1)$ 阶段,若别用 w 和 h 表示输入图像的宽和高,输入图像中所有坐标点(u,v)属于关键节点 k 的置信度值可表示为 $b_t^P\in\mathbf{R}^{w\times h}$,即

$$b_t^k(u,v)=b_t^k(Y_k=z) \tag{6}$$

交警手势包含 14 个关键节点,图像中交警手势所含所有关键节点的置信度集合表示为 $b_t\in\mathbf{R}^{w\times h\times 14}$.

通过上述步骤,可以为交警手势所含的每个部件产生置信度图. 经过 T 个阶段,置信度最高的位置即为关键节点位置. 即

$$Y_k=\text{argmax}(b_T^k),k\in\{1,\cdots,14\} \tag{7}$$

2.2 交警手势空间上下文特征提取

通过式(4)~(7)的计算可以确定交警手势中的每个关键节点的位置(如图 3 所示). 依据交警手势中骨架间的铰接依赖关系,可以通过相邻关键节点计算求得交警手势中骨架及其长度. 设 $\phi_1(\cdot)$ 为将部件位置转换为骨架矢量的函数. 即

$$\phi_1(Y_m,Y_n)\rightarrow s,s\in\boldsymbol{S} \tag{8}$$

本文使用骨架矢量提取了交警手势所包含的 2 种空间上下文特征 F_1、F_2. 其中,F_1 为骨架的相对可见长度;F_2 为骨架与重力方向的夹角,它们共同构成了交警手势的空间上下文特征集合 F. 即 $F=F_1\cup F_2$.

由于交警的头部长度为固定值,其不会随着身体的转动和摄像头距离的变化而改变. 因此,本文以交警头部长度为参考点,引入函数 $\phi_2(\cdot)$ 表示交警手势中所含骨架的相对可见长度的向量拼接,即

$$\phi_2(\boldsymbol{S})\rightarrow\oplus_{i=1}^{11}\frac{S_i}{\|\boldsymbol{S}_{\text{head}}\|},s_i\in\boldsymbol{S} \tag{9}$$

其中,$\boldsymbol{S}_{\text{head}}$ 是代表头顶至脖子中心的头部骨架矢量,$\|\cdot\|$ 表示矢量模,即头部骨架的长度. $\oplus$ 表示向量拼接. 该公式以 $\boldsymbol{S}_{\text{head}}$ 为参考,计算每个骨架相对于头部骨架的可见长度.

由于重力加速度的方向始终垂直于地面,为了描

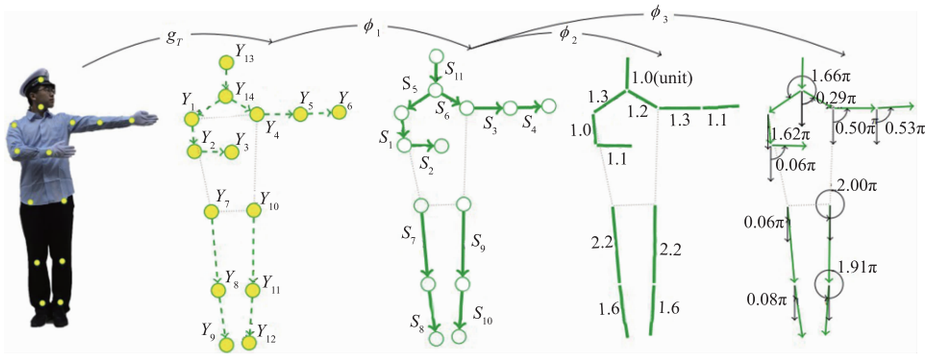


图3 交警手势空间特征提取过程示意

述交警手势中每个骨架相对于地面的方向,本文引入了骨架与重力加速度的夹角. 并使用 $\phi_3(\cdot)$ 表示每个骨架与重力方向夹角的向量拼接. 即

$$\phi_3(\boldsymbol{S})\rightarrow\left\{\oplus_{i=0}^{11}\frac{S_i\cdot\boldsymbol{d}}{\|S_i\|\ \|\boldsymbol{d}\|},\oplus_{i=0}^{11}\frac{S_i\times\boldsymbol{d}}{\|S_i\|\ \|\boldsymbol{d}\|}\right\} \tag{10}$$

为保持特征值的连续性,本文采用骨架与重力加速度方向的三角函数值来描述骨架的角度特征. 式(10)中,$\boldsymbol{d}$ 表示一个单位矢量,方向与重力方向相同. $\frac{S_i\cdot\boldsymbol{d}}{\|S_i\|\ \|\boldsymbol{d}\|}$ 计算了每个骨架矢量与重力方向夹角的 cos 值,$\frac{S_i\times\boldsymbol{d}}{\|S_i\|\ \|\boldsymbol{d}\|}$ 计算其 sin 值. 最终,由 $\phi_4(\cdot)$ 将上述两个特征拼接组合成为交警手势的空间上下文特征 F:

$$\phi_4(F_1,F_2)\rightarrow F_1\oplus F_2 \tag{11}$$

2.3　交警手势的时序特征提取

动态交警手势由一组具有时间先后顺序的图像序列组成. LSTM 是一种可以保持长时记忆的循环神经网络,可用于时间相关特征的提取. LSTM 依据式(12)保存记忆内容.

$$\begin{aligned}e_\tau=&\sigma(W_1\cdot(h_{\tau-1}\oplus F^\tau)+\beta_1)*e_{\tau-1}\\&+\sigma(W_2\cdot(h_{\tau-1}\oplus F^\tau)+\beta_2)\\&*\tanh(W_3\cdot(h_{\tau-1}\oplus F^\tau)+\beta_3)\end{aligned} \tag{12}$$

其中,h_τ 为输出的时间特征,e_τ 用于记忆保存,并作为下一个循环神经网络的输入. 在保存记忆的同时,LSTM 也依据式(13)计算输出向量 $\boldsymbol{h}_\tau$.

$$\boldsymbol{h}_\tau=\sigma(W_4\cdot(h_{\tau-1}\oplus F^\tau)+\beta_4)*\tanh(e_\tau) \tag{13}$$

其中,σ 为 sigmoid 函数,tanh 为 hyperbolic tangent 函数. $\oplus$表示向量拼接,$\cdot$ 表示矩阵乘法,$*$ 表示点乘. τ 代表当前时间. F^τ 表示在时间 τ 时的交警手势上下文空间特征. W 和 β 表示神经网络中可训练全连接层的权重和偏置.

最后,$\boldsymbol{h}_\tau$ 通过全连接层按照式(14)计算每类交警手势的预测概率,并按照式(15)将预测概率最大的手势作为预测手势.

$$o_\tau^{\text{all}}=s(W_5\cdot h_\tau+\beta_5) \tag{14}$$

$$o_\tau^{\max}=\begin{cases}\text{argmax}(o_\tau^{\text{all}}), & \max(o_\tau^{\text{all}})>\delta\\ 0, & \max(o_\tau^{\text{all}})<\delta\end{cases} \tag{15}$$

式(14)中,函数 $s(\cdot)$ 表示 softmax,o_τ^{all} 表示当前手势属于每个手势类的概率.

式(15)中,$o_\tau^{\max}$ 表示最终的手势分类输出. δ 表示动作置信度阈值,零值表示待机手势.

3　交警手势识别机

交警手势识别机 CTPGR 由关键节点提取网络、空间上下文特征提取器和 LSTM 网络构成,本节介绍它们的架构及相关网络训练方法.

3.1　交警手势关键节点提取网络

本文裁剪了 CPM 深度,并对输出中的每个关键节点进行前景背景二分类,交警关键节点提取网络 PKEN,图 4 所示为其网络架构.

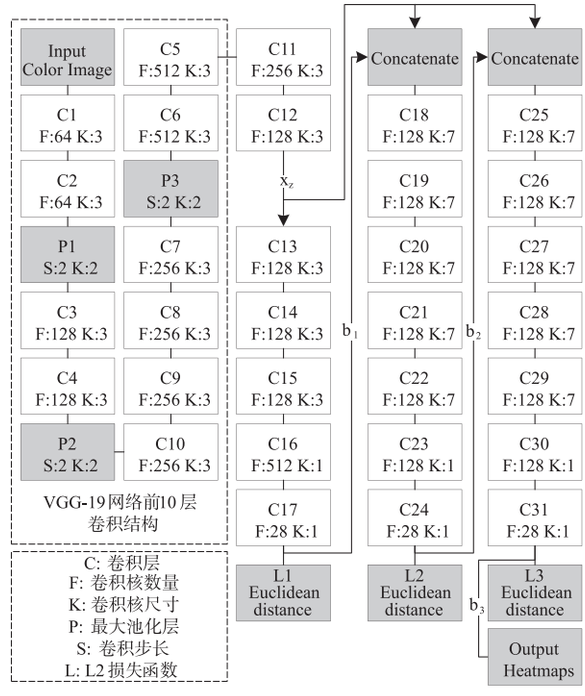


图4　PKEN网络架构

图 4 中,C 代表卷积层,P 代表最大池化层,L 代表 Loss 输出. 阶段数量 $t=3$. 从 C_1 至 C_{12} 的卷积网络实现了特征提取函数 $\psi(\cdot)$,即输出了图像中每个位置的特征值 x_z. C_{13} 至 C_{17} 层的卷积网络实现了第一阶段的分类器 $g_1(\cdot)$,它以 x_z 为输入,输出了交警手势中每个关键节点的置信度集合 $\boldsymbol{b}_1$. C_{18} 至 C_{24} 层的卷积网络实现了第二阶段的分类器 $g_2(\cdot)$,它以 x_z 和 $\boldsymbol{b}_1$ 为输入,输出了新感受野下交警手势中每个关键节点的置信度集合 $\boldsymbol{b}_2$. 最后,第 C_{18} 至 C_{24} 层的卷积网络实现了第三阶段的分类器 $g_3(\cdot)$,输出 $\boldsymbol{b}_3$. 图 5 所示为交警手势经 PKEN 处理输出的关键节点置信度图实例.

PKEN 一共包含了 3 个代价函数,分别是 L_1、L_2 和 L_3. 它们分别是 $\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 和 $\boldsymbol{b}_3$ 与真实的置信度之间的欧几里得距离. PKEN 的系统总 Loss 可按照式(16)计算出.

$$f=\sum_{j=1}^{14}\sum_{t=1}^{3}\sum_{z}\|b_t^j(z)-b_*^j(z)\| \tag{16}$$

式中,b_*^j 代表交警手势中第 j 个关键节点的真实置信度. z 代表置信度图中的像素集合.

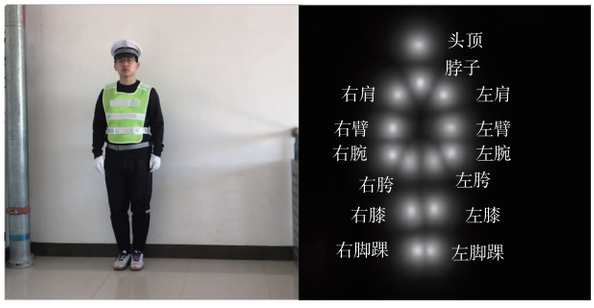


图5 交警手势关键节点置信度图

3.2 交警手势空间上下文特征提取

依据 PKEN 输出的关键节点及节点间的关联关系，根据式(9)和(10)可以分别计算出交警手势骨架的相对长度及其与重力加速度方向间的夹角，即生成在 τ 时刻的交警手势空间上下文特征 F_τ. 图 6 为计算手势空间上下文特征的伪代码，其中函数 PKEN 代表交警手势关键点提取网络，$\boldsymbol{Y}$ 代表其输出的置信度图. 根据节点间的关联关系 $\boldsymbol{S}$，程序计算骨骼矢量，并将被遮挡的骨骼赋予特殊矢量(0,0). 最后，计算骨骼长度、骨骼角度的 sin 值和 cos 值并收集在变量中.

```
Y = PKEN(image)
S = load_connection_dependency()
L_V = list() // connection vectors
for Y_m, Y_n in S:
    if Y_m < (0,0) or Y_n < (0,0):
        L_V.append(0,0)
    if Y_m > (0,0) and Y_n > (0,0):
        L_V.append(Y_n − Y_m)
L_F = list() //Features
for v in Lv:
    f = v / ‖v_head‖
    L_F.append(f)// bone length
    f = (S_i · d) / (‖S_i‖ ‖d‖)
    L_F.append(f)// bone angle:cos
    f = (S_i × d) / (‖S_i‖ ‖d‖)
    L_F.append(f) // bone angle:sin
return L_F
```

图 6 交警手势空间上下文提取伪代码

3.3 交警手势时序特征提取

LSTM 网络被用来提取动态交警手势的时间序列特征. 图 7 所示为本文所用 LSTM 网络的架构. 在图 7 中，$e_{\tau-1}$，$h_{\tau-1}$ 和 F_τ 是 LSTM 网络的输入. 其中，F_τ 是在 τ 时刻交警手势中各骨架的相对长度及其与重力加速度角的特征值. e_τ，和 h_τ 是网络的输出，并作为 $\tau>1$ 时 LSTM 网络的输入. 图中的“Dense”表示全连接层；P 表示逐点运算.

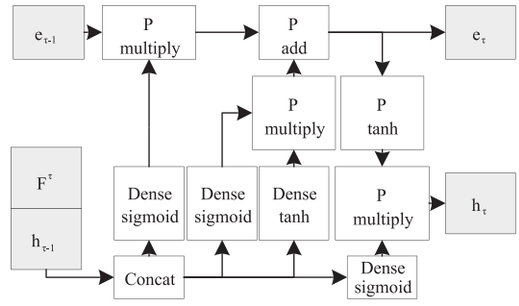


图7 LSTM架构

3.4 CTPGR 网络训练

如图 8 所示，CTPGR 网络训练包括人体关键节点网络 PKEN 训练、交警手势空间特征训练和 LSTM 时序特征训练 3 步. 作为数据集，本文按照中国交通交警手势规范录制了 8 种交警指挥手势视频共 21 段，如图 9 所示. 表 1 所示为样本集划分. 其中，“手势数量”指 8 种交警手势的合计样本数，“待机数量”指待机时段的数量.

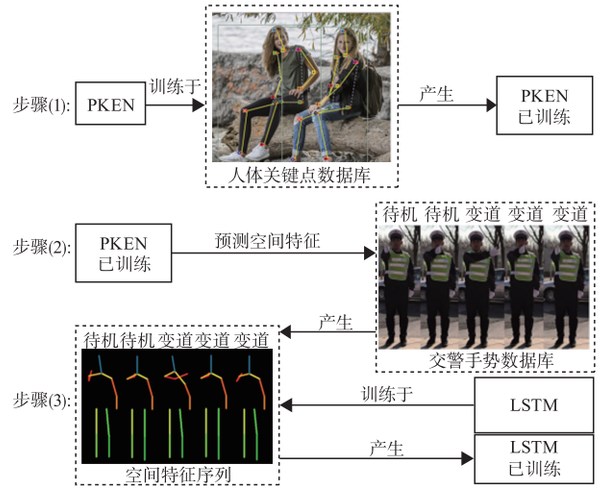


图8 CTPGR网络训练步骤

图9 交警手势数据集样本示例

表 1 交警手势数据集

集合	影片数量	手势数量	待机数量
训练集	11	1789	900
测试集	10	1565	787
总和	21	3354	1687

训练过程中，本文使用人体关键点数据[20]训练

PKEN,再使用 PKEN 网络从交警手势库中计算交警手势中骨架的相对长度和角度特征数据,与交警手势库中相应时刻标记的交警手势类型对应. LSTM 网络采用 Xavier 初始化[21],训练中对交警手势标签作 d 毫秒时延. 最后,将 PKEN、交警手势空间上下文特征提取器和 LSTM 网络依次连接构成了交警手势识别机 CTPGR.

4　实验与结果分析

为验证 CTPGR 在实际环境中的效果,课题组对其分别进行了离线测试和在线测试,度量标准为编辑距离. 同时,课题组将本文所提算法与已有的交警手势识别算法进行对比分析,证明本算法的有效性. 最后,课题组进行消融实验,通过移除各模块并计算准确率,证明了准确率的提升确实得益于本文提出的改进算法.

4.1　离线编辑距离实验

编辑距离(Edit Distance)被广泛用于评价两个序列之间的差异. 其直观意义为,求解序列 A 进行多少次变动能够使内容与序列 B 相同,变动包括插入、删除和替换. 编辑距离准确率如式(17)所示,其中,N 为视频中姿势总数,I 是视频中插入姿势的总数,D 是系统中删除姿势的总数,S 是系统中替换姿势的总数,C 是系统正确识别出的姿势总数.

$$准确率=\frac{(N-I-D-S)}{N} \tag{17}$$

离线实验步骤为,录制视频并标注,然后使用训练集数据训练网络,并预测测试数据集. 最后,将预测序列与标注序列进行编辑距离分析,结果如表 2 所示.

表 2　离线编辑距离

文件	预测	N	S	D	I	准确率
014	200	165	6	37	2	0.7273
016	186	161	2	27	2	0.8075
004	99	100	0	0	1	0.9900
018	165	163	1	4	2	0.9571
010	172	145	1	27	0	0.8069
104	161	167	0	0	6	0.9641
102	164	165	0	3	4	0.9576
012	179	177	0	2	0	0.9887
008	159	161	0	0	2	0.9876
002	170	161	0	9	0	0.9441
总数	1655	1565	10	109	19	0.9118

表 2 中,预测列表示视频中预测手势的总数,N 代表实际出现的手势总数. 其余列意义与式(17)中字母意义相同. 交警手势数据集的总编辑距离准确率为 91.18%,从表格分析得出,对于大部分室内、室外视频,即使光线不足、背景复杂且有车辆运动、交警位置有变化,本模型依然能够保持较高的准确率. 但是,本方法在繁华马路表现较差,还需通过更多工作区分交警与行人.

4.2　在线编辑距离实验

本文使用编辑距离对模型进行了实时在线测试,其难点在于模型运行的速度能够影响最终结果. 在线测试视频共持续 5 分钟,表 3 为在线实时测试编辑距离准确率,系统的准确率达到 93.82%,与离线测试结果相持平.

表 3　在线系统测试结果

N	I	D	S	C	准确率
81	2	1	2	78	93.82

4.3　时延长度与准确率关系

本文使用离线编辑距离在数据集上对时延 d 和准确率的关系进行定量实验,结果如图 10 所示.

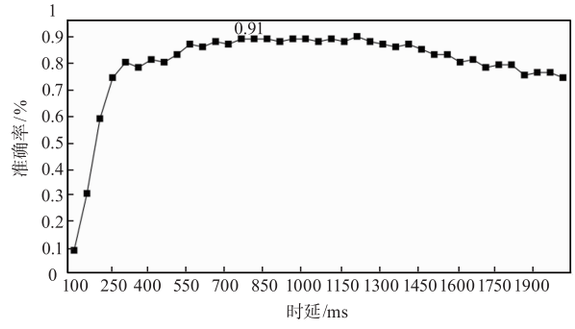


图10　时延长度与准确率关系

图 10 中,横轴代表训练时时延的长度,纵轴代表使用该时延长度训练的模型在测试集上的准确率. 实验以 50ms 为间隔绘制准确率随时延的变化趋势. 从图中可以看出,在 250ms 前,准确率随时延增加而迅速增加;在 250ms 至 750ms 范围中,准确率缓慢增加,在 750ms 附近到达顶峰. 本文选取了 750ms 作为方法的时延值.

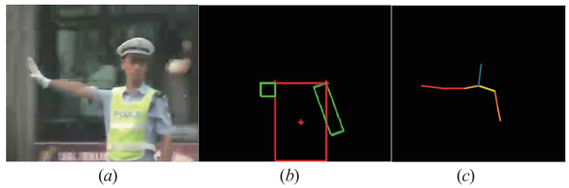
4.4　CTPGR 与其它算法的对比分析

本文实现了参考文献[22~26]中的算法,并将它们与本算法进行对比,结果如表 4 所示. 表 4 中前四种方法都通过提取人体骨架数据进行姿势识别. 其中,文献[22]采用双向 LSTM 网络,其准确率与本文基本持平,但其网络不能用于在线实时分析. 文献[23]和文献[24]与本文采用了不同的关键点定位网络,其中最高准确率达到了 89%. 文献[25]采用端到端的 3D 卷积网络识别交警手势,其识别率为 81.46%;文献[26]采用卷积 LSTM 网络识别交警手势,其识别率为 82.40%. 表 4 表明:本文采用 PKEN 提取交警手势的空间上下文特征,并结合 LSTM 网络提取交警手势的时序特征,是一种有效的交警手势识别方法.

表 4 不同姿势识别算法对比

No.	Architecture	Reference	Accuracy
1	(Ours) PKEN + LSTM		91.18%
2	PKEN + Bidirectional LSTM	L Pigou, et. al. [22]	91.04%
3	Resnet 定位部件 + LSTM	He, et. al. [23]	87.22%
4	Densenet 定位部件 + LSTM	Huang, et. al. [24]	89.66%
5	End to end 3D Convolution	Ji, et. al. [25]	81.02%
6	Convolutional LSTM	Xing, et. al. [26]	80.77%

本文采用 PKEN 提取交警手势部件的空间上下文信息,相比核密度估计[27]等方法,本方法需要更多训练数据,但具有更好的抗噪声干扰能力. 以图 11 为例. 其中,图 11(*a*)为被检测的交警手势原图像;图 11(*b*)为采用文献[27]的算法检测交警手势部件的输出结果,绿色长方形代表交警的手臂、红色长方形代表躯干;图 11(*c*)为本算法检测交警部件的输出结果,其包含了交警上半身头顶、脖子、左右肩膀、左右肘和左右手腕共 8 个部件的位置. 将原图 11(*a*)与图 11(*b*)、图 11(*c*)对比可看出,原图中交警的右手向身体右侧伸直,但图 11(*b*)输出的结果中右臂方向垂直于纸面,只能看见一个方块. 这是因为这张图片背景混杂、噪声很大,导致文献[27]的算法出现了误判. 在图 11(*c*)中,橘红色和红色的线段代表了交警的右手小臂和大臂. 图 11(*c*)表明本文的算法能够正确识别交警右臂所处的位置,说明本算法相比文献[27]具有更强的抗干扰能力. 另外,图 11(*c*)相比图 11(*b*)包含了更多交警手势关键节点的信息,对手势具有更精确的描述能力.



(*a*) (*b*) (*c*)

图11 关键节点提取网络的优势

4.5 消融实验

CTPGR 分别采用 PKEN 提取交警手势的关键节点,手势空间特征提取器提取交警手势中骨架的长度及其与重力加速度方向的夹角,LSTM 提取交警手势的时序特征. 课题组分别移除 CTPGR 中的手势空间特征提取器和 LSTM 构造了两种交警手势识别算法,进行了对比实验. 表 5 为仅由 PKEN 和 LSTM 构成的交警手势识别算法与 CTPGR 的手势识别准确率对比.

从表 5 可以看到,对于大部分的测试视频,CTPGR 由于增加了手势空间特征提取过程使得其手势识别准确率有了提升,表明手势空间上下文特征提取是本文 CTPGR 中非常重要的组成部分.

表 5 由 PKEN 和 LSTM 构成的手势识别算法与 CTPGR 的准确率对比

文件名	*N*	*S*	*D*	*I*	由 PKEN 和 LSTM 构成的手势识别算法准确率	CTPGR 的识别准确率
014	165	13	33	5	0.6909	0.7273
002	161	1	4	4	0.9441	0.9441
004	100	1	5	1	0.9300	0.9900
018	163	0	1	4	0.9693	0.9571
010	**145**	**17**	**36**	**0**	**0.6345**	**0.8069**
012	177	6	12	2	0.8870	0.9887
104	167	0	0	6	0.9641	0.9641
016	161	2	26	1	0.8198	0.8075
102	165	0	0	6	0.9636	0.9576
008	161	0	0	0	0.9804	0.9876
TOTAL	1565	40	117	29	0.8811	0.9118

此外,课题组保留 CTPGR 中的 PKEN 和手势空间特征提取模块,而去除了 LSTM 模块,导致手势识别网络失去了记忆能力,输出了许多错误的手势类别,说明时间特征记忆不可缺少. 表 6 显示了此方法对 10 个交警手势视频文件的测试结果.

表 6 无 LSTM 模块交警手势识别结果

文件名	实际手势数量	预测手势数量	错误预测手势数量
018	163	1247	1084
010	145	820	675
014	165	1659	1494
016	161	1517	1356
002	161	1029	868
012	177	1058	881
104	167	858	691
102	165	931	766
008	161	974	813
004	100	660	560

4.6 CTPGR 速度测试

课题选取了具有驾照的 8 名年龄 21 ~ 25 岁自愿者,通过采用相同的交警指挥手势视频来对比被试和 CTPGR 识别交警手势的平均响应时间. 方法最高帧率为 17.2fps,可稳定输出 15fps. 在实验中,被试要求观看交警指挥手势视频,在清晰辨认出交警做出的每个手势后立刻按下确认键,用于仿真计算驾驶员识别交警手势的响应时间. 图 12 所示为 CTPGR 和被试识别 8 种交警手势的平均响应时间对比. 图 12 中,横轴代表不同类别的手势;纵轴为时间长度. 图 12 中绿色矩形代表被

试识别交警手势的平均响应时间;橘黄色矩形代表实验中每个手势的标签出现的延迟时间,固定为 750ms. 蓝色矩形和橘黄色矩形合并高度代表 CTPGR 识别手势的响应时间. 从图 12 可以看出,被试手势识别的响应间比 CTPGR 方法更短,但 CTPGR 方法的响应时间标准差更小. 总体而言,个体和 CTPGR 的手势识别响应时间差距不大(最大 0.28s);此外,在识别"左转"手势时被试的平均响应时间(1.1s)大于本 CTPGR 的平均响应时间(0.93s). 在真实交通场景中,个体通常在完整地看清交警手势后再做出识别,其响应时间通常在 2s 左右. 在图 12 中,本文 CTPGR 方法的手势识别的平均响应时间在 1.38s 之内,属于比较合理的范围.

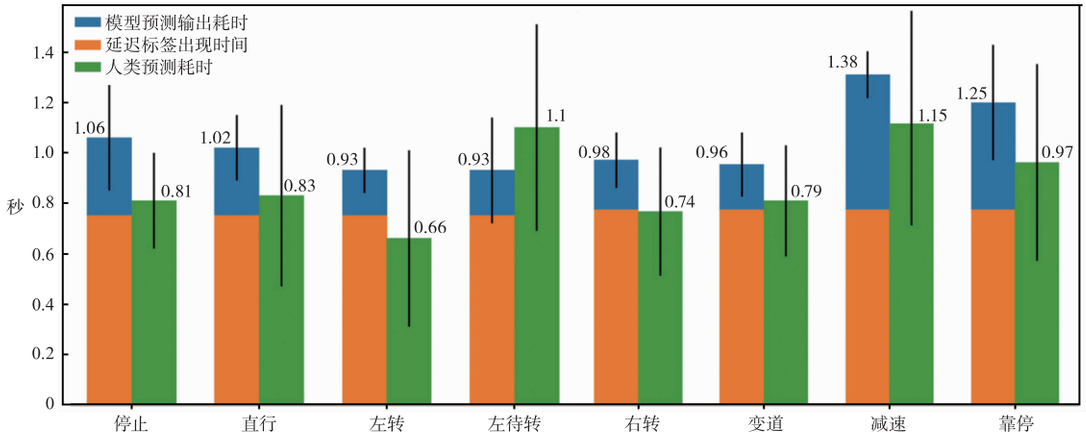


图12　八类手势识别反应时间

5　结论

本文将 CPM 扩展为 PKEN 网络,并和 LSTM 网络相结合,提取交警手势的关键节点,并在此基础上提取交警手势骨架的相对长度和与重力加速度夹角作为交警手势上下文空间特征;最后将这些空间上下文特征作为 LSTM 网络的输入,提炼交警手势的时序特征. 实验证明本文提出的交警手势识别方法可以有效的提炼交警手势的时空特征,能适应背景复杂以及交警手势动态变化等应用场景,系统的准确率达到 91.18%,同时系统的响应时间较短,可以满足交警手势实时识别的需要. 本文的 CTPGR 系统目前只支持视频中含有单个交警的情况,课题组将借鉴参考文献[18]的思想,通过添加交警着装等特征设计支持多人场景的交警手势识别系统.

参考文献

[1] Wang B, Yuan T. Traffic police gesture recognition using accelerometer[A]. IEEE SENSORS Conference[C]. Lecce-Italy, 2008. 1080 – 1083.

[2] Yuan T, Wang B. Accelerometer-based Chinese traffic police gesture recognition system[J]. Chinese Journal of Electronics, 2010, 19(2): 270 – 274.

[3] 杜友田,陈峰,徐文立,等. 基于视觉的人的运动识别综述[J]. 电子学报,2007,35(1):84 – 90.
DU You-tian, CHEN Feng, XU Wen-li, et al. A survey on the vision-based human motion recognition[J]. Acta Electronica Sinica, 2007, 35(1): 84 – 90. (in Chinese)

[4] Cai Z, Guo F. Max-covering scheme for gesture recognition of Chinese traffic police[J]. Pattern Analysis and Applications, 2015, 18(2): 403 – 418.

[5] Guo F, Tang J, Cai Z. Automatic recognition of Chinese traffic police gesture based on max-covering scheme[J]. Advances in Information Sciences and Service Sciences, 2013, 5(1): 428.

[6] Eichner M, Ferrari V. Human pose co-estimation and applications[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 34(11): 2282 – 2288.

[7] Eichner M, Marin-Jimenez M, Zisserman A, et al. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images[J]. International Journal of Computer Vision, 2012, 99(2): 190 – 214.

[8] 管业鹏. 复杂人机交互场景下的指势用户对象识别[J]. 电子学报,2014,42(11):2135 – 2141.
GUAN Ye-peng. Pointing user recognition in human-computer interaction with cluttered scene[J]. Acta Electronica Sinica, 2014, 42(11): 2135 – 2141. (in Chinese)

[9] Kang H, Lee C W, Jung K. Recognition-based gesture spotting in video games[J]. Pattern Recognition Letters, 2004, 25(15): 1701 – 1714.

[10] 张友梅,常发亮,刘洪彬. 基于 3D 人体骨架的动作识别[J]. 电子学报,2017,45(4):906 – 911.
ZHANG You-mei, CHANG Fa-liang, LIU Hong-bin. Action recognition based on 3D skeleton[J]. Acta Electronica Sinica, 2017, 45(4): 906 – 911. (in Chinese)

[11] Zhang Z. Microsoft kinect sensor and its effect[J]. IEEE Multimedia,2012,19(2):4 – 10.

[12] Guo F,Tang J,Wang X. Gesture recognition of traffic police based on static and dynamic descriptor fusion[J]. Multimedia Tools and Applications,2017,76(6):8915 – 8936.

[13] Le Q K,Pham C H,Le T H. Road traffic control gesture recognition using depth images[J]. IEIE Transactions on Smart Processing & Computing,2012,1(1):1 – 7.

[14] Zhou Z,Li S,Sun B. Extreme learning machine based hand posture recognition in color-depth image[A]. Chinese Conference on Pattern Recognition[C]. Springer,2014. 276 – 285.

[15] 罗会兰,童康,孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报,2019,47(5):1162 – 1173.
LUO Hui-lan,TONG Kang,KONG Fan-sheng. The progress of human action recognition in videos based on deep learning:a review[J]. Acta Electronica Sinica,2019,47(5):1162 – 1173. (in Chinese)

[16] Fragkiadaki K,Levine S,Felsen P,et al. Recurrent network models for human dynamics[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. US:IEEE,2015. 4346 – 4354.

[17] Wei S-E,Ramakrishna V,Kanade T,et al. Convolutional pose machines[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. US:IEEE,2016. 4724 – 4732.

[18] Cao Z,Simon T,Wei S-E,et al. Realtime multi-person 2d pose estimation using part affinity fields[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. US:IEEE,2017. 7291 – 7299.

[19] Ramakrishna V,Munoz D,Hebert M,et al. Pose machines:Articulated pose estimation via inference machines[A]. European Conference on Computer Vision[C]. Berlin:Springer,2014. 33 – 47.

[20] Wu J,Zheng H,Zhao B,et al. Large-scale datasets for going deeper in image understanding[A]. 2019 IEEE International Conference on Multimedia and Expo (ICME)[C]. US:IEEE,2019. 1480 – 1485.

[21] Glorot X,Bengio Y. Understanding the difficulty of training deep feedforward neural networks[A]. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics[C]. US:ACM,2010. 249 – 256.

[22] 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 – 4):430 – 439.

[23] He K,Zhang X,Ren S,et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. US:IEEE,2016. 770 – 778.

[24] Huang G,Liu Z,Van Der Maaten L,et al. Densely connected convolutional networks[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. US:IEEE,2017. 4700 – 4708.

[25] Ji S,Xu W,Yang M,et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(1):221 – 231.

[26] Xingjian S H I,Chen Z,Wang H,et al. Convolutional LSTM network:A machine learning approach for precipitation nowcasting[A]. Advances in Neural Information Processing Systems[C]. CSDN,2015. 802 – 810.

[27] Guo F,Cai Z,Tang J. Chinese traffic police gesture recognition in complex scene[A]. IEEE 10th International Conference on Trust,Security and Privacy in Computing and Communications[C]. US:IEEE,2011. 1505 – 1511.

作者简介

张 丞 男,1993 年 11 月出生,北京人. 分别于 2016 年和 2019 年在北京工业大学获得学士、硕士学位. 现为北京工业大学博士研究生,主要研究方向为智能人机交互和模式识别.

何 坚(通讯作者) 男,1969 年 12 月出生,新疆阿拉尔人. 2000 年获得西安西北大学硕士学位,2005 年获得西安交通大学博士学位. 现为北京工业大学软件学院副教授,主要研究方向为智能人机交互、普适计算和物联网.
E-mail:jianhee@bjut.edu.cn