时空特征与通道注意力融合的视觉手势识别技术

何 坚[1,2], 刘 炎[2], 祖天奇[2]
(1. 北京市物联网软件与系统工程技术研究中心, 北京 100124; 2. 北京工业大学信息学部, 北京 100124)

摘 要: 为了解决双流融合网络对动态手势关键帧及手部轮廓特征检测不足的问题,提出一种手势时空特征与通道注意力融合的动态手势识别方法. 首先,在双流融合网络中引入有效通道注意力(eficient channel attention,ECA)增强双流识别算法对手势关键帧的关注度,并利用双流中的空间卷积网络和时间卷积网络分别提取动态手势中的空间和时序特征;其次,通过 ECA 在空间流中选取最高关注度的手势帧,利用单发多框检测器技术(single shot multibox detector,SSD)提取相应手部轮廓特征;最后,将手部轮廓特征与双流中提取的人体姿态特征、时序特征融合后分类识别手势. 该方法在 Chalearn 2013 多模态手语识别数据集上进行了验证,准确率为 66.23%,相比之前在该数据集上仅使用 RGB 信息进行双流识别的方法获得了更好的手势识别效果.

关键词: 动态手势识别; 双流融合网络; 通道注意力; 关键帧; 单发多框检测器; Chalearn 2013 数据集
中图分类号: TP 391 **文献标志码**: A **文章编号**: 0254 - 0037(2021)08 - 0824 - 09
doi: 10.11936/bjutxb2020120028

Visual Gesture Recognition Based on Spatial-Temporal Features and Channel Attention

HE Jian[1,2], LIU Yan[2], ZU Tianqi[2]
(1. Beijing Engineering Research Center for IoT Software and Systems, Beijing 100124, China;
2. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: To solve problems of insufficient detection of dynamic gesture key frames and hand contour features in two-stream fusion network, a dynamic gesture recognition method was proposed in this paper based on the fusion of spatial-temporal features and channel attention. First, the efficient channel attention (ECA) was introduced into the two-stream fusion network to enhance the attention of key frames of gestures, the spatial convolutional network and the temporal convolutional network of two-stream were used to extract spatial and temporal features of dynamic gestures. Second, the gesture frame with the highest attention in the spatial network was selected by ECA, and single shot multibox detector (SSD) was used to extract the hand contour features. Finally, hand contour features were integrated with body posture features and temporal features were extracted from two-stream to recognize gestures. The method proposed in this paper was verified on Chalearn 2013 multi-modal sign language recognition dataset, with an accuracy rate of 66.23%. Compared with the previous two-stream methods which only RGB information from this dataset was adopted, it achieves a better gesture recognition effect.

Key words: dynamic gesture recognition; two-stream fusion; channel attention; key frames; single shot multibox detector; Chalearn 2013 dataset

收稿日期: 2020-12-28
基金项目: 国家自然科学基金资助项目(61602016);国家重点研发计划资助项目(2020YFB2104400)
作者简介: 何 坚(1969—), 男, 副教授, 主要从事物联网、移动普适计算技术及人机交互技术方面的研究, E-mail:Jianhee@bjut.edu.cn

近年来，手势识别技术在体感游戏、手语识别、辅助驾驶及智能家电控制等领域应用广泛. 由于手势在人机交互中的重要性，手势识别系统的研究一直是人们关注的焦点. 根据文献[1]的调查，自然手势的表达大多是动态的，通过人体手部和上半肢协调运动来完成. 因此，动态手势相比静态手势在手势表达中更为重要.

文献[2-3]中对近年来的一些动态手势识别方法进行了总结. 例如，Adewuyi 等[4]结合手指和手臂肌肉的肌电图数据对手部抓握和手指动作进行分类；Huang 等[5]通过双通道方法融合人体手部加速度、角速度及肌肉电数据，再结合 K 邻近（K-nearest neighber，KNN）算法识别手势；田元等[6]使用 Kinect 体感设备获取人体的骨骼信息和深度图信息，结合骨骼关节点位置及手指特征对手势进行实时识别.

上面提到的一些工作为了达到更好的识别效果使用了不止一种模态信息，这种情况称为多模态方法[7]. 它们通常结合颜色信息（RGB 格式）、深度图信息和骨骼关节点信息来检测识别手势. 这种多源信息，如深度图和骨架关节，积极补充了颜色信息，有助于手势分类[8]，然而除 RGB 以外信息的获取，通常需要特定的传感器，如微软的 Kinect、华硕 Xtion Pro 或英特尔的 Realsense3. 这种对特定传感器的依赖导致对交互环境的限制，影响手势的自然表达. 相反，基于 RGB 视频数据的动态手势识别技术具有使用方便、成本较低等优点，另外在许多公共空间也很容易找到监控摄像机，交互环境更多. 这也是促使人们致力于发展仅使用 RGB 视频数据识别动态手势的原因之一.

之前的研究工作中，也有些工作仅将 RGB 视频数据作为手势识别的唯一信息来源，但只有少数工作者取得到了显著的结果，如文献[8-9]. 即便如此，这些研究中提到的好的识别效果也是在固定的几类手势动作上实现的，这些手势在表达时身体动作差异较大，通常也会简化手势识别的任务.

最近，深度学习的一些方法在计算机视觉领域的几个问题上分别取得了最优的结果[9-10]，该类方法通常使用三维卷积网络（3D convolutional neural networks，3DCNN）[11]、双流融合网络[12-13]、卷积神经网络（convolutional neural network，CNN）和长短时记忆（long short term memory，LSTM）网络组合的方式[14]来识别动态手势. 例如，Nunez 等[15]通过 CNN 从连续手势帧中提取人体的骨骼数据和手部骨骼数据，再结合 LSTM 识别动态手势；Al-Hammadi 等[16]直接使用 3DCNN 识别动态手势；Zhang 等[17]将 3DCNN 和 LSTM 结合从视频帧中学习手势的时空特征图，再利用 CNN 从该特征图中学习更高层次的时空特征用于手势识别.

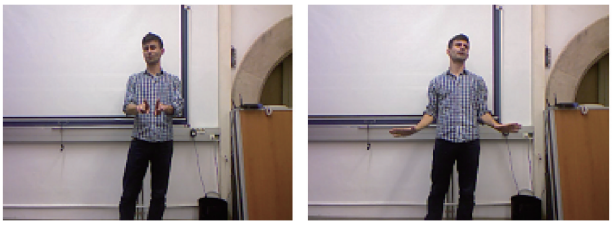
文献[13]中双流融合的方法在 HMDB51[18]和 UCF101[19]2 个人体动作数据集上取得最佳识别效果. 该方法通过 2 个卷积网络分别提取连续人体动作的空间特征和时序特征（光流），并探讨对比不同光流提取算法及双流融合方法对人体动作识别效果的影响，证明双向光流能较好表达人体运动信息. 不过该方法应用于手势识别任务仍存在 2 个主要缺点：1）未对不同时序帧的初始权重系数进行考虑；2）空间通道直接对整幅视频帧卷积操作，对较小手部特征关注度不足.

最近的一些研究发现注意力机制能够帮助深度学习从众多信息中抽选出对当前任务目标更为关键的信息[20]，其核心思想是基于原有数据找到数据间的关联性，进而突出某些重要特征. 而有效通道注意力（efficient channel attention，ECA）[21]机制相比同类型注意力机制降低了模型的复杂度并获得更高准确度. 受双流融合网络和 ECA 注意力机制启发，本文对双流融合网络进行改进，结合有效通道注意力机制和单发多框检测器技术（single shot multibox detector，SSD）[22]建立了基于视觉的动态手势识别模型，并在 Chalearn 2013 公开手势数据集[23]上进行实验验证.

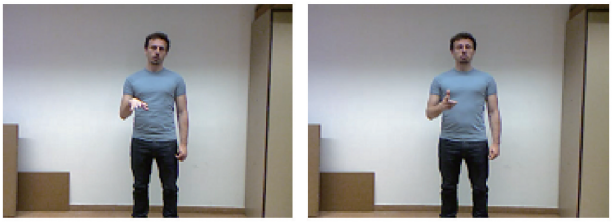
1　动态手势建模

手势交互环境中，动态手势的形态主要由人体姿态及手部轮廓构成，连续性手势的表达涉及对其变化规律的考虑. 双流融合网络的方法分别从空间和时间上提取手势特征，对身体姿态差别较大的手势识别较好，但对身体姿态相同手部具体形状不同的手势识别欠佳，如图 1 所示（图 1 中（a）（b）2 个手势在表达时身体动作差别较大；（c）（d）2 个手势身体动作相同但手部轮廓不同）. 分析原因是因为双流融合网络在空间流中直接对整幅手势图像卷积操作，对较为明显的人体姿态特征能够有效提取，但对局部较小的手部轮廓特征关注不足.

本文首先在双流融合网络中引入 ECA 注意力机制增强双流识别算法对手势关键帧的关注度；其次选取关注度最高的手势帧提取手部轮廓特征；最后将补充的手部轮廓特征与双流特征融合后分类识别手势.



(a) 合不来 (b) 不够



(c) 放在一起 (b) 没有了

图 1 不同手势举例

Fig. 1 Examples of different gestures

表 1 中汇总了本文主要特征和有关映射函数的数学符号表示.

表 1 主要符号和含义对应表

Table 1 Main symbols and associated meanings

符号	含义
X	平均选取的手势帧集合
I	光流图集合
G	人体姿态特征
O	手部轮廓特征
S	手势运动特征
F	融合后的手势特征
C	通道数目
k	ECA 中的跨通道交互数目
$\sigma(\cdot)$	Sigmoid 激活函数
$g(\cdot)$	全局平均池化
$\phi(\cdot)$	k 与 C 的映射函数
$\psi(\cdot)$	三维卷积 + 三维池化
$R(\cdot)$	转换为一维向量

1.1 手势双流特征提取

相比静态手势,动态手势的识别还需要考虑连续帧之间的手势动作变换规律. 光流法是利用图像序列中像素在时间域上的变化以及相邻帧之间的相关性来计算出相邻帧之间人体运动信息的一种方法[24]. 另外,利用光流作为时序上的运动信息可以去除不同背景对手势识别的影响. 本部分参考双流融合网络的思想建立了动态手势双流卷积网络(gesture two-stream convolution network, GTSCN),该网络分别从空间和时间上提取手势表达中的人体姿态特征、运动特征,结构如图 2 所示.

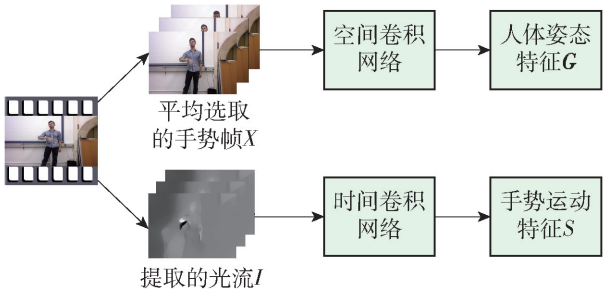


图 2 GTSCN 网络结构

Fig. 2 GTSCN network structure

对于一个输入宽、高分别为 w、h 的手势视频,首先平均选取 T 帧手势图 $X_\tau, X_{2\tau}, \cdots, X_{T\tau}$,将其堆叠作为双流中空间卷积网络的输入,用来提取动态手势中的人体姿态特征 G. 其中每选取的 2 帧手势图之间相隔 τ 帧.

对于手势的运动特征则通过双流中的时间卷积网络来提取. 时间流的输入由连续手势帧之间叠加的光流位移场构成. 参考文献[25],稠密的光流可以看作一系列沿运动轨迹位移场向量 $\boldsymbol{d}_t$ 的集合,其中 $\boldsymbol{d}_t$ 表示第 t 和 $t+1$ 连续帧之间的位移场向量,$\boldsymbol{d}_t$ 的水平和垂直分量 $\boldsymbol{d}_t^x$ 和 $\boldsymbol{d}_t^y$ 分别表示相邻帧之间的手势沿不同方向的运动轨迹.

针对选取 T 帧手势中的任一帧 X_τ,本文设计将其左右各 $L/2$ 帧的光流图叠加,形成总共 $2L$ 个通道的双向光流 $\boldsymbol{I}_\tau^{w\times h\times 2L}$ 作为双流中时间卷积网络的输入,用来提取手势的运动特征 S. I_τ 的严格定义为

$$\begin{cases} I_\tau(u,v,2k-1)=\boldsymbol{d}^x_{\tau+k+1}(P_k) \\ I_\tau(u,v,2k)=\boldsymbol{d}^y_{\tau+k+1}(P_k), u\in[1,w], \\ \quad v\in[1,h], k\in[1,L] \end{cases} \tag{1}$$

式中 P_k 表示从第 τ 帧 (u,v) 位置开始,沿着这个轨迹的第 k 个点,用来记录手势在每一帧像素上的移动轨迹,并且有如下递推解释:

$$\begin{cases} P_1=(u,v) \\ P_k=P_{k-1}+\boldsymbol{d}_{\tau+k-2}(P_{k-1}), \ k>1 \end{cases} \tag{2}$$

1.2 注意力机制关键帧选取

以上建立的双流卷积网络模型分别从堆叠的手势帧和光流帧中提取手势特征. 需要注意的是动态手势的表达是一个时序过程,注重手势表达过程中易于区分的关键性动作更能增强手势的识别效果.

本文引入 ECA 通道注意力对输入双流卷积网

络中的手势帧和光流帧的特征图通道集合进行加权，用来提升手势关键帧的关注度.

ECA 的工作原理在于：通过学习每个特征图通道在整个特征图通道集合中的权重比例系数，进而增强权重较高特征图通道的学习. 通过将手势帧和光流帧按照时间顺序堆叠（每个手势帧和光流帧都可以看作一个特征图通道），再结合 ECA 即可求取每个手势帧和光流帧的加权权重，权重最高的即为动态手势表达过程中的关键帧.

另外，由于时序上堆叠的特征图通道之间具有一定的局部周期性[26]（时间间隔较远视频帧之间的相关性更小），假设每个视频帧对应的特征图通道仅与其邻近 k 个特征图通道相关，依据 ECA 注意力机制的思想可结合每个特征图通道的邻近 k 个通道计算出该通道的局部加权权重

$$\begin{cases} w_i = \sigma\left(\sum_{j=1}^{k} \alpha^j y_i^j\right) \\ y_i = g(c_i),\ c_i \in C \end{cases} \tag{3}$$

式中：C 表示需要加权的特征图通道集合；c_i 表示 C 中的第 i 层特征图通道；σ 表示 Sigmoid 激活函数；函数 $g(\cdot)$ 表示全局平均池化. 设 k 与 C 之间的映射关系为 $\phi(\cdot)$，依据 ECA 本文使用以下非线性函数映射 $\phi(\cdot)$

$$\begin{cases} C = \phi(k) = 2^{\gamma k - b} \\ k = \left| \dfrac{\mathrm{lb}C}{\gamma} + \dfrac{b}{\gamma} \right|_{\mathrm{odd}} \end{cases} \tag{4}$$

式中：$|\mathrm{num}|_{\mathrm{odd}}$ 表示将 num 向上舍入为最近的奇数；γ、b 为任意自然系数，本文设 $\gamma = 2$、$b = 1$. 至此，识别动态手势关键帧的注意力机制模块已建立. 将输入 GTSCN 网络中的手势帧集合 X 和光流帧集合 I 分别代入到式(3)中的 C，即可求对应通道的加权权重，进而增强手势关键帧的识别.

1.3　手部轮廓特征选取

通过式(3)可计算出 GTSCN 空间卷积网络中每一帧手势的加权权重，由于手势表达过程中的一些特定手部形态可以帮助区分不同手势，因此本文选取加权权重最高的手势帧用来提取手部轮廓特征 O，增强运动姿态相似但手部轮廓不同手势的识别效果.

这里只选择加权权重最高的手势帧提取手部轮廓特征的考虑是：动态手势的表达是一个时序过程，手势表达过程中的初始阶段和结束阶段包含信息不多，如果对每一帧的手势都提取手部轮廓特征，作用性不强也增加计算复杂度. 因此本文设计只提取关键帧的手部轮廓特征.

其中手部轮廓 O 共包含 O_l、O_r 两部分，分别表示手势关键帧中左右手预测为不同手部轮廓类别的置信度集合，如 O_l^i 表示左手属于第 i 类手部轮廓的置信度. O_l、O_r 中置信度最高的即为对应的左右手类别. 在此基础上，将左右手轮廓特征 O 与 GTSCN 网络中提取的人体姿态特征 G、手势运动特征 S 融合即可构成动态手势的时空上下文特征 F.

值得注意的是，GTSCN 网络中提取的人体姿态特征 G 和运动特征 S 具有像素级的对应关系. 以刷牙和梳头 2 个动作为例，如果一只手在某个空间位置周期性地移动，那么时间卷积网络就能识别其运动轨迹，而空间卷积网络就可以识别其形态（牙齿或毛发），将其组合就可以辨别动作. 因此本文首先在通道维度上堆叠特征 G、S 用来满足特征图层的像素级对应关系，然后使用三维卷积和三维池化对特征 G、S 进行融合，最后设计在全连接层拼接手部轮廓特征 O，有

$$F = R(\psi(G \oplus S)) \oplus O \tag{5}$$

式中：$\oplus$ 表示变量拼接或通道堆叠；$\psi(\cdot)$ 表示对变量进行三维卷积和三维池化；$R(\cdot)$ 表示将变量转换为一维变量. 最后 F 通过全连接层即可计算每类手势的预测概率 p_i，预测概率最大 $p^{\max}$ 即可作为最终的预测手势.

2　动态手势识别机

本文建立的动态手势识别机如图 3 所示，由通道注意力模块、手部轮廓特征提取网络、双流卷积网络、特征融合及分类模块构成，其中双流卷积网络中的空间卷积网络和时间卷积网络均均采用 VGG16[27] 构建，其他部分的构建方法将在本节逐一介绍.

2.1　通道注意力模块

本文选用 ECA 来构建通道注意力模块，图 4 即为 ECA 结构的示意图. 对于输入通道为 C 的手势帧和光流帧，首先使用全局平均池化操作（golbal average pooling，GAP）将每一层的特征图通道 c_i 都映射为一个单一变量 l_i. 这里使用该操作的原因是：为保障求取各个视频帧对应特征图通道权重系数的合法性，应该结合整个通道的空间上下文信息，另外将整个特征图通道映射为单一变量也可以减少网络的训练参数，进而降低模型复杂度. 全局平均池化操作的工作原理为：利用当前特征图通道中所有位置像素值的平均值用来表达整个特征图通道的信息.

其次，需要结合每个特征图通道对应变量的

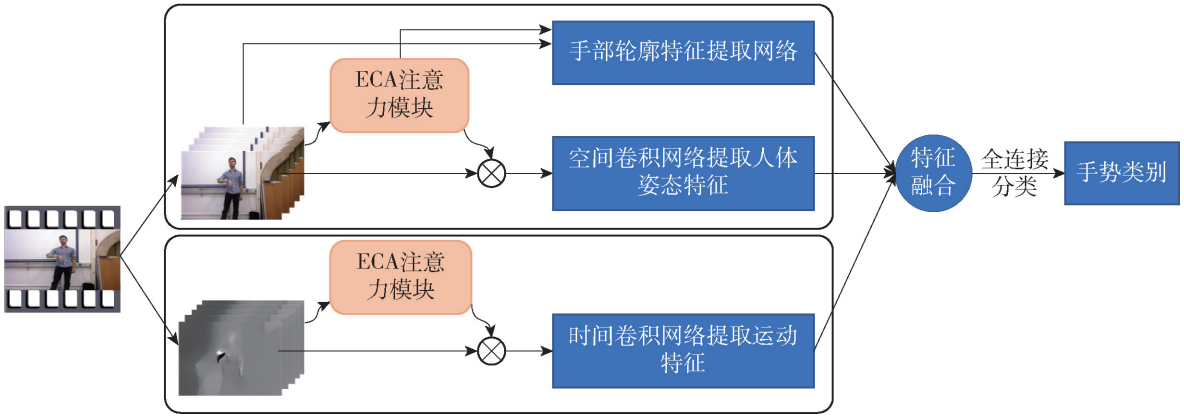


图 3 动态手势识别框架

Fig. 3 Dynamic gesture recognition frame

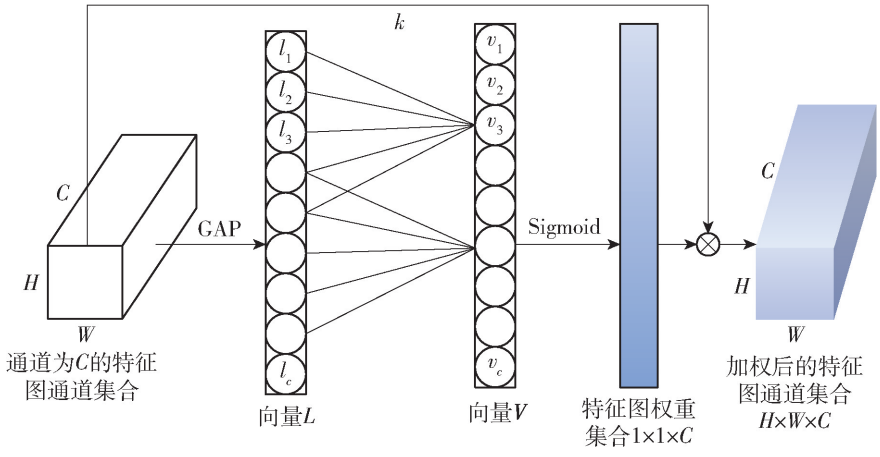


图 4 注意力机制模块图

Fig. 4 Attention module diagram

邻近 k 个变量计算出当前特征图通道的加权权重 v_i(时序上堆叠的特征图通道具有一定的局部周期性,邻近的特征图通道之间相互影响,可只计算每个通道在其邻域内的加权权重). v_i 的计算表达式为

$$v_i = \sum_{j=1}^{k} \alpha_i^j l_i^j = \alpha_i^1 l_i^1 + \alpha_i^2 l_i^2 + \cdots + \alpha_i^k l_i^k \tag{6}$$

式中 l_i^j 表示第 i 个特征图通道邻近的第 j 个通道的输出值.

最后,使用 Sigmoid 激活函数将每个特征图通道的权重归一化到[0,1]范围内再结合输入数据即可得到加权后的特征图通道集合. ⊗表示乘法操作.

其中,向量 $\boldsymbol{L}$ 和向量 $\boldsymbol{V}$ 之间的映射矩阵为

$$\begin{bmatrix}
w^{1,1} & w^{1,2} & w^{1,3} & \cdots & w^{1,k} & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & w^{2,2} & w^{2,3} & \cdots & w^{2,k} & w^{2,k+1} & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & w^{3,3} & \cdots & w^{3,k} & w^{3,k+1} & w^{3,k+2} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & w^{c-1,c-1} & w^{c-1,c} & \cdots & w^{c-1,c+k-2} & 0 \\
0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & w^{c,c} & \cdots & w^{c,c+k-2} & w^{c,c+k-1}
\end{bmatrix} \tag{7}$$

式中:第 i 行第 k 个非零元素 $w^{i,k}$ 即表示第 i 个特征图通道邻近的第 k 个特征图通道对其影响权重.

2.2 手势轮廓特征提取网络

输入空间卷积网络中的连续手势帧经过 ECA 模块可选出手势表达过程中的关键帧,然后再利用手势

轮廓特征提取网络可从关键帧中提取手部轮廓特征，用来弥补双流融合网络对较小手部轮廓检测不足的问题. 需要注意的是，多生物特征融合虽然可以提高识别系统的准确性，但也无疑提高了计算复杂度.

SSD 作为一种多尺度、高精度的目标检测技术，能够快速识别图片中物体的位置及类别，因此本部分引用 SSD 技术从手势关键帧中提取手部轮廓特征，图 5 即为本文所用 SSD 网络架构. 其中卷积层 conv_8 ~ conv_11 分别从不同尺度的特征图中提取手部轮廓进行分类，旨在解决不同用户的不同手部大小对手部轮廓分类的影响. 该方法的具体实现思路是：首先，在多个不同尺度特征图层的每个像素点周围预设几个候选框；然后，针对每个候选框都预测距离真正手部位置的偏移量及各类手部轮廓的置信度；最后，选择偏移量较小候选框中置信度最高的类别作为最终的手部轮廓类别.

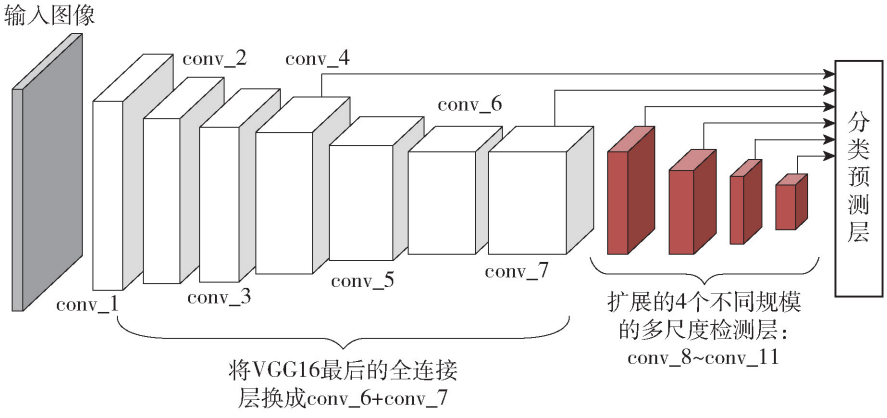


图 5　手部轮廓特征提取网络结构图

Fig. 5　Network structure diagram of hand contour feature extraction

2.3　特征融合分类模块

本文建立的特征融合及手势分类结构如图 6 所示，其中 $\psi(\cdot)$ 表示对变量进行三维卷积和三维池化操作，$R(\cdot)$ 表示将变量转换为一维变量，⊕表示变量拼接或通道堆叠，FC 表示全连接层. 对于 GTSCN 双流网络中提取的人体姿态特征图 G 和手势运动特征图 S，首先，在通道维度上进行堆叠并使用三维卷积和三维池化操作融合特征 G、S 形成动态手势的双流特征 D；其次，将融合后的双流特征 D 转换成一维变量，并与手部轮廓特征 O 进行变量拼接；最后，使用全连接层进行分类得到最终的手势识别类型.

其中，全连接层的作用在于：通过多次线性变换求取融合后的手势特征 F 属于每一类手势类型的概率，概率最高的即为最终的手势类型. 另外，本文的损失函数定义为交叉熵损失函数（全连接层的多

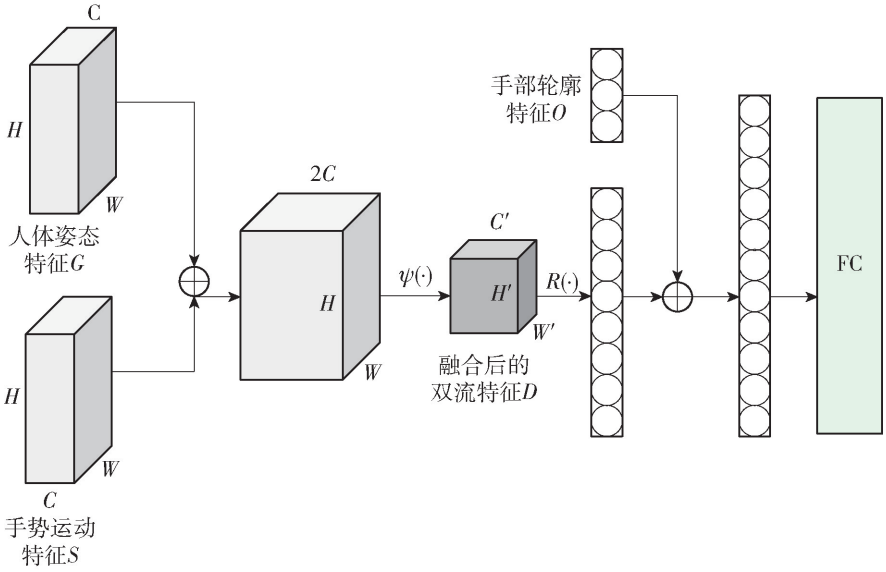


图 6　特征融合分类模块示意图

Fig. 6　Feature fusion and classification module diagram

分类任务常使用该损失函数),即

$$H(X) = -\frac{1}{M}\sum_{i=1}^{M} p_i \ln(p_i) + (1-p_i)\ln(1-p_i) \tag{8}$$

式中:M 表示手势类型个数;p_i 表示手势属于第 i 个类型的概率.

3 实验结果分析

3.1 数据集简介

为验证本方法的泛化性,本文选择公开的 Chalearn 2013 意大利手语数据集进行实验. 该数据集使用 Kinect 传感器以每秒 20 帧的速度记录了 27 个用户在不同背景下的手势词汇表达,其中共包含 20 个手势分类,每个手势的时长在 50 帧左右,并提供 RGB、RGB-D、骨架、用户轮廓多种模态信息. 另外该数据集共计 13 858 个样本,其中训练集 7 754 个、验证集 3 362 个、测试集 2 742 个. 本文使用该数据的 RGB 模态数据与其他仅使用 RGB 信息的动态手势识别方法进行了对比.

3.2 网络训练

3.2.1 双流结构

本文设计 GTSCN 网络中的空间和时间卷积网络均采用 VGG16 特征提取网络构建,包含 5 个卷积层和 3 个全连接层,有关 VGG16 的具体参数设置可参考文献[27].

由于 Chalearn 2013 视频数据的分辨率为 640 × 480,因此对于空间卷积网络,首先按照手势样本的开始帧和结束帧在中间平均选取 T 帧;然后从这 T 帧手势图中随机剪裁 480 × 480 的区域并缩放到 224 × 224 的分辨率大小;最后将堆叠的维度为 224 × 224 × T 的手势帧输入到空间卷积网络.

对于时间卷积网络,首先按照式(1)计算出选取 T 帧手势图中每一帧手势的光流图集合,然后将堆叠的维度为 224 × 224 × 2L × T 的光流图输入到时间卷积网络.

3.2.2 手部轮廓特征提取网络

本文截取 Chalearn 2013 视频数据的手势帧标注左右手候选框及对应手部轮廓类型,进而训练手部轮廓特征提取网络. 具体实现步骤如下:

步骤 1 在 38 × 38 的 conv_4、19 × 19 的 conv_7、10 × 10 的 conv_8、5 × 5 的 conv_9、3 × 3 的 conv_10、1 × 1 的 conv_11 六个不同尺度特征图层中预设多个手部标记候选框. 其中,每一个候选框都需要预测以下 2 点信息:候选框中的手部轮廓类型 p;左下角及右上角 2 个顶点坐标($x_{\min}$,$y_{\min}$)、($x_{\max}$,$y_{\max}$)距离真正手部位置的偏移量.

步骤 2 针对这 6 个特征图层中的每一个候选框,都使用 5 个卷积过滤器利用卷积操作的方式得到预测的 4 个坐标偏移量及手部轮廓类型置信度.

步骤 3 将各个候选框中预测的手部轮廓类型置信度从大到小排序,选取置信度最高的候选框作为其中一个手的真实框,并将其预测的手部轮廓类型和位置坐标作为该手部的预测结果.

步骤 4 计算剩余候选框与当前真实框的重叠度(intersection over union, IOU),并根据预设的重叠度阈值 IOU_t 过滤掉一部分候选框(若上一步已确定左手的真实框,则可以过滤掉剩余所有左手的候选框,本文设置 IOU_t 为 0.5). 然后从剩余候选框中选择预测置信度最高的作为另外一个手的真实框. 重叠度的计算公式为

$$\mathrm{IOU} = \frac{\mathrm{box}_1 \cap \mathrm{box}_2}{\mathrm{box}_1 \cup \mathrm{box}_2} = \frac{\mathrm{box}_1 \cap \mathrm{box}_2}{\mathrm{box}_1 + \mathrm{box}_2 - (\mathrm{box}_1 \cap \mathrm{box}_2)} \tag{9}$$

式中 box_i 表示第 i 个候选框的面积.

另外,由于 Chalearn2013 手势数据集中的左右手轮廓在视频图片中占比较小,因此本文按照式(10)对 SSD 默认候选框的归一化尺度做了调整,即

$$\mathrm{scale} = \frac{\mathrm{BoxSize}}{\mathrm{ImageSize}} \tag{10}$$

实验时标注的左右手候选框大小与视频画面的尺度比多数为 0.05 ~ 0.30,因此本文设计手部轮廓候选框的归一化尺度见表 2.

表 2 手部轮廓候选框的归一化尺度

Table 2 Normalized scale of the hand contour candidate box

方法	1	2	3	4	5	6
SSD	0.20	0.34	0.48	0.62	0.76	0.90
本文	0.05	0.10	0.15	0.20	0.25	0.30

3.2.3 特征融合及分类模块

本文设计的特征融合及手势分类结构如图 6 所示. 其中,三维卷积核的维度为 3 × 3 × 3,步长为 1;三维池化的维度为 2 × 2 × 2,步长为 2(最大池化). 另外,本文在全连接层后面添加 softmax 激活函数预测手势类别.

3.3 实验结果分析

在显卡为 NVIDIA Titan X、处理器为 Intel Xeon ES 的实验环境下,本文方法与之前在该数据集上的

最佳手势识别结果进行了对比，如表 3 中所示.

文献[28]中设计的耦合隐式马尔可夫算法(coupled hidden Markov model，CHMM)仅使用 RGB 信息在该数据集上获得了之前的最佳手势识别效果，准确率为 60.07%. 该方法通过集成 2 个或多个隐式马尔可夫链(HMM)学习不同链隐藏节点的相互作用，进而增强单 HMM 的识别效果. 本文一开始设计的 GTSCN 网络结构分别从空间卷积网络和时间卷积网络中获取手势的时空上下文信息，实验准确率为 64.57%；结合 ECA 注意力机制后实验准确率为 65.84%；再通过补充 SSD 提取的手部轮廓特征后获得了 66.23% 的识别效果.

由上述分析可知，通过结合通道注意力和手部轮廓特征，可有效提高双流融合网络的手势识别准确率.

表 3　不同方法的实验结果对比

Table 3　Results of different methods

序号	方法	准确率/%
1	CHMM	60.07
2	GTSCN	64.57
3	GTSCN + ECA	65.84
4	本文方法	66.23

本文也实验对比了不同特征融合策略对手势识别结果的影响，如表 4 所示. Max 方法表示选取特征 G、S 在相同空间位置特征图通道的最大值作为双流融合特征，Sum 方法表示选取特征 G、S 在相同空间位置特征图通道的和作为双流融合特征. 实验结果表明利用三维卷积和三维池化能够更好地提高手势识别准确率，分析原因是与二维卷积、二维池化相比，三维卷积、三维池化能更好地从视频序列中学习人体手势的运动变化规律，其卷积和池化操作都是在时空上执行，而二维卷积和二维池化仅在空间上完成.

表 4　不同特征融合策略对实验结果影响对比

Table 4　Effect of different feature fusion strategies on experimental results

序号	方法	准确率/%
1	Max	63.15
2	Sum	65.79
3	本文方法	66.23

另外，在本文的实验环境中，SSD 识别关键帧中的手部轮廓类型约耗时 50 ms，相邻两帧之间的光流计算约耗时 11 ms(光流在视频播放过程中采取实时计算)，识别的总体延迟时间在 200 ms 以内，因此本文的手势识别方法可基本满足实时性要求.

4　结论

1) 提出了一种基于 RGB 视频数据的动态手势识别模型. 首先依据双流融合网络的思想构造了 GTSCN 网络，用来提取动态手势中的人体姿态特征、运动特征；其次设计在 GTSCN 网络中引入 ECA 注意力增强手势关键帧的学习，并结合 SSD 提取手部轮廓特征；最后通过全连接层分类识别手势.

2) 通过在 Chalearn 2013 公开手语数据集上进行实验，证明结合 ECA 和 SSD 可以增强双流算法对相似手势的识别效果.

3) 下一步研究计划是针对本文提出的模型开发设计一个手势识别系统，将实时拍摄到的手势视频转换为对应文本含义.

参考文献：

[1] LIU H, WANG L. Gesture recognition for human-robot collaboration: a review [J]. International Journal of Industrial Ergonomics, 2018, 68: 355-367.

[2] XIA Z, LEI Q, YANG Y, et al. Vision-based hand gesture recognition for human-robot collaboration: a survey [C] // 2019 5th International Conference on Control, Automation and Robotics (ICCAR). Piscataway: IEEE, 2019: 198-205.

[3] PATIL N M, PATIL S R. Review on real-time EMG acquisition and hand gesture recognition system[C]//2017 International conference of Electronics, Communication and Aerospace Technology (ICECA). Piscataway: IEEE, 2017: 694-696.

[4] ADEWUYI A A, HARGROVE L J, KUIKEN T A. An analysis of intrinsic and extrinsic hand muscle EMG for improved pattern recognition control [J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2015, 24(4): 485-494.

[5] HUANG D, ZHANG X, SAPONAS T S, et al. Leveraging dual-observable input for fine-grained thumb interaction using forearm EMG [C] // The 28th Annual ACM Symposium. New York: ACM, 2015: 523-528.

[6] 田元，王学璠，王志锋，等. 基于 Kinect 的实时手势识别方法[J]. 计算机工程与设计，2018，39(6)：1721-1726.
TIAN Y, WANG X P, WANG Z F, et al. Real-time hand

gesture recognition method using Kinect[J]. Computer Engineering and Design, 2018, 39(6): 1721-1726. (in Chinese)

[7] NEVEROVA N, WOLF C, TAYLOR G, et al. Moddrop: adaptive multi-modal gesture recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(8): 1692-1706.

[8] ESCALERA S, ATHITSOS V, GUYON I. Challenges in multi-modal gesture recognition[J]. Journal of Machine Learning Research, 2016, 17: 1-54.

[9] DARGAN S, KUMAR M, AYYAGARI M R, et al. A survey of deep learning and its applications: a new paradigm to machine learning [J]. Archives of Computational Methods in Engineering, 2020, 27(4): 1071-1092.

[10] LIU W, WANG Z, LIU X, et al. A survey of deep neural network architectures and their applications[J]. Neurocomputing, 2017, 234: 11-26.

[11] 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.

[12] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1: 568-576.

[13] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 1933-1941.

[14] 张丞, 何坚, 王伟东. 空间上下文与时序特征融合的交警指挥手势识别技术[J]. 电子学报, 2019, 48(5): 966-974.
ZHANG C, HE J, WANG W D. Visual recognition of Chinese traffic police gestures based on spatial context and temporal features [J]. Acta Electronica Sinica, 2019, 48(5): 966-974. (in Chinese)

[15] NUNEZ J C, CABIDO R, PANTRIGO J J, et al. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition[J]. Pattern Recognition, 2018, 76: 80-94.

[16] AL-HAMMADI M, MUHAMMAD G, ABDUL W, et al. Hand gesture recognition using 3D-CNN model[J]. IEEE Consumer Electronics Magazine, 2019, 9(1): 95-101.

[17] ZHANG L, ZHU G, SHEN P, et al. Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition [C]// International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE, 2017: 3120-3128.

[18] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]// International Conference on Computer Vision. Piscataway: IEEE, 2011: 2556-2563.

[19] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2020-12-28]. https://arxiv.org/abs/1212.0402.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[21] WANG Q, WU B, ZHU P, et al. ECA-net: efficient channel attention for deep convolutional neural networks [EB/OL]. [2020-12-20]. https://arxiv.org/abs/1910.03151.

[22] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]// European conference on computer vision. Berlin: Springer, 2016: 21-37.

[23] ESCALERA S, GONZALEZ J, BARO X, et al. Multi-modal gesture recognition challenge 2013: dataset and results[C]// Association for Computing Machinery. New York: ACM, 2013: 445-452.

[24] AGARWAL A, GUPTA S, SINGH D K. Review of optical flow technique for moving object detection[C]// International Conference on Contemporary Computing and Informatics. Piscataway: IEEE, 2016: 409-413.

[25] 李元祥, 谢林柏. 基于深度运动图和密集轨迹的行为识别算法[J]. 计算机工程与应用, 2020, 56(3): 194-200.
LI Y X, XIE L B. Human action recognition based on depth motion map and dense trajectory[J]. Computer Engineering and Applications, 2020, 56(3): 194-200. (in Chinese)

[26] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting [J]. Advances in Neural Information Processing Systems, 2015, 1: 802-810.

[27] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020-12-28]. https://arxiv.org/abs/1409.1556.

[28] CAO C, ZHANG Y, LU H. Multi-modal learning for gesture recognition[C]// IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2015: 1-6.

(责任编辑 郑筱梅)