



## 时域注意力 Dense-TCNs 在多模手势识别中的应用

张 毅, 赵杰煜, 王 翀, 郑 烨

(宁波大学 信息科学与工程学院, 浙江 宁波 315211)

**摘 要:** 为增强时间卷积网络(TCNs)在时间特征提取方面的能力, 提出一种基于三维密集卷积网络与改进 TCNs 的多模态手势识别方法。通过时空特征表示方法将手势视频分析任务分为空间分析和时间分析两部分。在空间分析中采用三维 DenseNets 学习短期的时空特征, 在时间分析中使用 TCNs 提取时间特征。在此基础上引入注意力机制, 使用时域维度的压缩-激励网络调整每个 TCNs 层特征在时间维度上的权值比重。分别在 VIVA 和 NVGesture 两个动态手势数据集上对该方法进行评价, 实验结果表明, 该方法在 VIVA 数据集上的正确率为 91.54%, 在 NVGesture 数据集上的正确率为 86.37%, 且与最新的 MTUT 方法水平相近。

**关键词:** 手势识别; 三维密集卷积网络; 时间卷积网络; 短时时空特征; 注意力机制

开放科学(资源服务)标志码(OSID):



**中文引用格式:** 张毅, 赵杰煜, 王翀, 等. 时域注意力 Dense-TCNs 在多模手势识别中的应用[J]. 计算机工程, 2020, 46(9): 101-109.

**英文引用格式:** ZHANG Yi, ZHAO Jieyu, WANG Chong, et al. Application of time domain attention Dense-TCNs in multimodal gesture recognition[J]. Computer Engineering, 2020, 46(9): 101-109.

## Application of Time Domain Attention Dense-TCNs in Multimodal Gesture Recognition

ZHANG Yi, ZHAO Jieyu, WANG Chong, ZHENG Ye

(College of Information Science and Engineering, Ningbo University, Ningbo, Zhejiang 315211, China)

**[Abstract]** In order to enhance the temporal feature extraction ability of Temporal Convolutional Networks(TCNs), this paper proposes a multimodal gesture recognition method based on 3D Dense convolutional Networks(3D-DenseNets) and improved TCNs. 3D-DenseNets are used in spatial analysis to effectively learn short-term temporal and spatial features, and TCNs are used to extract temporal features in temporal analysis. On this basis, the attention mechanism is introduced, and the time-domain compression-stimulation network is used to adjust the weight ratio of each TCN layer feature in the time dimension. The method is evaluated on two dynamic gesture data sets, VIVA and NVGesture. Experimental results show that the proposed method achieves an accuracy rate of 91.54% on VIVA and 86.37% on the benchmark of NVGesture, reaching a level similar to that of the latest MTUT method.

**[Key words]** gesture recognition; 3D Dense convolutional Networks(3D-DenseNets); Temporal Convolutional Networks(TCNs); short-term temporal and spatial features; attention mechanism

**DOI:** 10.19678/j.issn.1000-3428.0056808

### 0 概述

随着科学技术的迅猛发展, 手势识别已成为当前科学研究领域的热点之一, 其主要应用领域有目标检测<sup>[1]</sup>、视频检索、人机交互<sup>[2]</sup>、手语识别<sup>[3]</sup>等。由于手势识别存在相似手势之间的细微差别、复杂的场景背景、不同的观测条件以及采集过程中的噪

声等, 使得通过机器学习得到一个鲁棒性手势识别模型具有较大的挑战性。

基于深度学习的手势识别主要任务是从图像或视频中提取特征, 然后将每个样本分类或确定到某个标签上。手势识别旨在识别和理解手臂与手在其中起着关键作用的人体有意义的运动。但是在动态手势视频中, 一般只有少量的手势可以从图像或单

**基金项目:** 国家自然科学基金(61603202, 61571247); 浙江省自然科学基金重点项目(LZ16F03001, LY17F030002)。

**作者简介:** 张 毅(1994—), 男, 硕士研究生, 主研方向为计算机视觉; 赵杰煜, 教授; 王 翀, 副教授; 郑 烨, 硕士研究生。

**收稿日期:** 2019-12-05 **修回日期:** 2020-03-26 **E-mail:** rhettzhang@live.com

个视频帧中的空间或结构信息中识别出来。事实上,运动线索和结构信息同时表征了一个独特的手势,而如何有效地学习手势的时空特征一直是手势识别的关键。尽管在过去的几十年中,人们提出了很多方法来解决这个问题,如从静态手势到动态手势,从基于运动轮廓到基于卷积神经网络,但是在识别精度方面仍然存在不足。

目前,现有的基于深度学习的孤立手势识别模型已经拥有了较高的识别率,多数方法都是基于卷积神经网络(CNNs)<sup>[4-5]</sup>或递归神经网络(RNNs)<sup>[6]</sup>开发的。

随着深度学习的发展,越来越多新颖且高效的网络体系结构被提出,其中比较有代表性的方法为文献[7]提出的密集卷积神经网络(DenseNets),相比于传统的CNNs, DenseNets拥有更深的网络层级结构,并且模块内的卷积层互相密集关联,从而使网络在拥有深层次结构的同时,避免由于网络过深而导致信息丢失的问题。实验结果表明, DenseNets拥有较高的特征提取能力和识别率。而针对复杂的手势,三维CNNs能够有效地学习到视频内连续视频帧中的手势短时的空间、结构和姿态变换,这是单帧图像或图片的二维CNNs所欠缺的。但由于在传统三维CNNs模型训练过程中,作为输入的视频片段(较短的连续帧)会有重复输入的部分出现,且如果重复部分较大则会大幅延长模型的训练速度,因此如何简化学习操作与高效训练模型是一个十分重要的课题。

对于时序模型而言,文献[8]提出一种新的解决序列问题的结构——时间卷积神经网络(TCNs)。与传统的RNNs及其典型的递归体系结构LSTMs和GRUs相比,TCNs具有较好的清晰性和简单性。

为提取更完整更有代表性的特征信息,文献[9]证明了在神经网络中特征信息的内部存在多种关系,并提出将注意力机制作为深度学习模型的嵌入模块。而压缩-激励网络 SENets<sup>[10]</sup>是一个高效的基于注意力机制的体系结构单元,其目标是通过显式地建模其卷积特征通道之间的相互依赖性来提高网络生成的质量表示。

本文采用三维 DenseNets 提取多段基于连续视频帧片段的短时空特征,并组成一条由短时空特征组成的序列。将短时空特征序列输入到TCNs中完成分类任务,并采用针对时间维度改进的压缩-激励方法(TSE),增强TCNs在时间特征提取方面的能力。

## 1 相关研究

基于视觉的手势识别技术包括面向静态手势的方法和面向动态手势的方法<sup>[2]</sup>。近年来,CNNs<sup>[4]</sup>凭借其强大的特征提取能力,在计算机视觉相关任务

上取得了重大突破,因此,CNNs提取的特征被广泛应用于许多动作分类任务中以获得更好的性能。二维卷积网络(2D-CNNs)最初是应用于二维图像中的,也就是静态手势或者是动态手势视频中的单帧图像,如文献[11-12]使用二维CNN并通过多层等级池化对图像手势进行识别,提取空间与时域上的信息。而三维卷积网络(3D-CNNs)的发展,使得三维卷积(C3D)在后续的研究中被广泛应用。文献[13]将三维CNNs引入到动态视频手势识别中,具有较好的性能,该研究的主要贡献是提出了一种从视频片段中提取时空特征的体系结构。另一方面,文献[14]设计了一个用于手势识别的多流3D-CNNs分类器,该分类器由两个子网络组成:高分辨率网络(HRN)和低分辨率网络(LRN),这为后续研究提供了宝贵经验。为解决视频中的手势片段训练的问题,文献[15]提出了一种新的时间池化方法。

随着深度卷积神经网络的发展,越来越多的CNNs体系结构被提出,如 AlexNets<sup>[11]</sup>、VGGNets<sup>[16]</sup>、GoogleNets<sup>[17-20]</sup>、ResNets<sup>[21]</sup>和 DenseNets<sup>[7]</sup>。上述模型的目标就是构建一个更高层次的CNNs体系结构,从低层次的图像帧中挖掘更深入、更完整的统计特征,然后进行分类。在孤立手势识别领域,文献[22]使用 Res-C3D 模型应用于手势识别任务中。文献[23]同样适用 Res-C3D 模型,并在2016年和2017年的ChaLearn LAP多模态孤立手势识别挑战赛<sup>[24-25]</sup>中两次获得第一名,这足以证明层次越深的网络拥有更强的特征学习能力。而 DenseNets<sup>[7]</sup>作为最新的卷积结构之一,逐渐被应用于动作识别,特别是人脸识别<sup>[26]</sup>和手势识别。除图像识别领域外,在最近的研究中, DenseNets 也被用来对不同的行为进行分类,如文献[27]使用 DenseNets 进行行为识别的研究。而深度信息作为除RGB信息外的额外视频信息被国内外研究所应用,其中文献[28-29]使用深度图对手势进行识别。

对于视频序列的时间信息,LSTM网络是手势识别的常用选择。例如,文献[30]将卷积长短期记忆模型(conv-LSTM)引入到时空特征图中,从而通过手势视频中的前后关系进行识别。文献[31]使用2S-RNN(RGB和深度图)进行连续手势识别。然而,包括LSTMs和GRUs在内的RNNs在时域上存在着短时信息学习、存储容量过大等缺点。为了弥补这些不足,人们提出TCNs并将其应用于手势再现中。文献[32]提出了基于骨架的动态手势识别方法 Res-TCNs,实验结果表明,相较于传统的RNNs,TCNs在结构上更加简洁,并能够有效地提高识别率。在整个时序数据中有许多冗余信息,因此,引入注意力机制显得非常重要。文献[33-34]在使用时序模型的同时,嵌入了相关的注意力机制模型,在原有时序模型识别率的基础上降低了错误率。

本文方法具体工作如下:

1) 为解决单帧图像不能承载足够的手势空间和结构信息,而多视频帧训练又需要避免视频片段所导致数据重复训练的问题,结合截断的 3D-DenseNets (T3D-Dense) 和局部时间平均池化 (LTAP) 两种方法作为短时空特征序列的提取模型。

2) 利用时间卷积网络代替传统的递归神经网络作为短时空特征序列分析的主要模型,并对压缩-激励网络 (SENet) 进行改进,使其能够应用于时域维度嵌入 TCNs 中,重新调整层间的短时空特征序列的权值,从而更有效地对短时空特征序列进行分析,达到更高的分类精度。

## 2 时域注意力 Dense-TCNs

本文提出一种新的模型来提取时空特征,并对时空特征序列进行识别和分类。模型流程如图 1 所示,整个过程可分为以下 2 个部分:

1) 通过截断的 T3D-Dense、局部时间平均池 (LTAP) 和多模式特征串接提取多模式的短时空特征序列模块。

2) 基于 TCN 和 TSE 的时空特征序列识别模块。

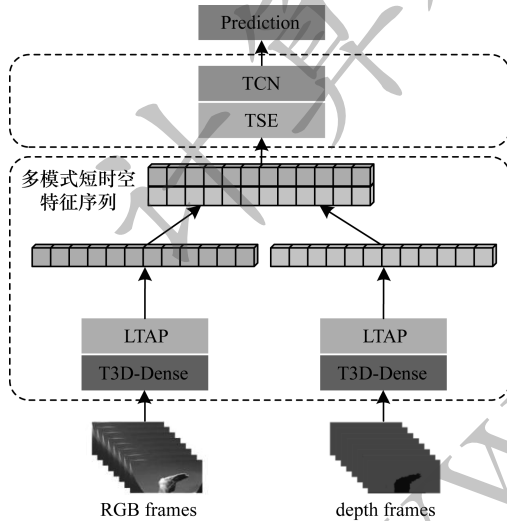


图 1 本文模型的流程  
Fig. 1 Procedure of the proposed model

### 2.1 短时空特征学习

由于签名视频的性质,一个健壮的视频特征表示需要结合多模态手势信息。在手势视频中,前后帧之间存在着多种关系,包括位置、形状和序列信息。因此,本文设计一个基于 C3D 的多流 DenseNets 作为时空特征提取器,从视频中提取时空特征。在此模型中,所有视频集的长度必须相同。因此,一个给定的视频  $V$  和  $n$  帧需要规范化为  $k$  帧。本文设置的输入为:

$$V_s = [V_1, V_2, \dots, V_k] \quad (1)$$

其中,  $V_k$  是输入视频序列的第  $k$  帧图像。

如前文所述,本文考虑多种形式的手势视频数据作为输入。每种类型的数据被设置为一个数据流并馈送到相同的网络结构,它们的输出随后将被融合,见图 1。每个数据流共享相同的网络结构,网络结构如表 1 所示,模型包含 4 个致密块体,每个区块包含 6、12、24、16 层密集连接卷积层,网络的增长率为 12,表 1 中显示的每个“conv”层对应于 BN-ReLU-conv 序列。值得注意的是,大多数卷积层都使用  $3 \times 3 \times 3$  大小的卷积核,同时在空间和时域上进行分析,但是为了避免短期时间信息的融合,将所有过渡层的时间池大小和步长设置为 1,这主要是区别与其他传统的 C3D 模型。

表 1 三维 DenseNets 架构  
Table 1 3D-DenseNets architecture

层	过滤器大小
Convolution	$5 \times 5 \times 5$ conv, stride $2 \times 2 \times 1$
Pooling	$3 \times 3 \times 1$ max pool, stride $2 \times 2 \times 1$
Dense Block 1	$\begin{bmatrix} 1 \times 1 \times 1 \text{ conv} \\ 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer 1	$1 \times 1 \times 1 \text{ conv}$ $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block 2	$\begin{bmatrix} 1 \times 1 \times 1 \text{ conv} \\ 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer 2	$1 \times 1 \times 1 \text{ conv}$ $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block 3	$\begin{bmatrix} 1 \times 1 \times 1 \text{ conv} \\ 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Transition Layer 3	$1 \times 1 \times 1 \text{ conv}$ $2 \times 2 \times 1$ average pool, stride $2 \times 2 \times 1$
Dense Block final	$\begin{bmatrix} 1 \times 1 \times 1 \text{ conv} \\ 3 \times 3 \times 3 \text{ conv} \end{bmatrix} \times 16$
Classification Layer	global spatial average pool global temporal average pool fully-connected, softmax

由于三维 DenseNets 是一个短期的时空特征提取器,因此在本文中其被截断,使模型只得到经过全局空间平均池化后的特征。具体来讲,首先用孤立的手势数据对模型进行预训练,然后丢弃全局时间平均池层、最后一个 softmax 层和完全连接层。因此,模型可以在全局空间平均池层之后得到全局时空特征  $F_k$ :

$$F_k = [f_1, f_2, \dots, f_k] \quad (2)$$

其中,时间长度为  $k$ ,并且表示  $k$  帧的短时空特征。

然后从全局特征  $F_k$  中剪切和合并  $T$  个短时空特征。第  $t$  个短期时空特征  $x_t$  构造为:

$$x_t = \text{ltap}[f_{t-\frac{k}{T}}, f_{t-\frac{k}{T}+1}, \dots, f_{t+\frac{k}{T}-1}] \quad (3)$$

其中,ltap 是截断的 3D-DenseNets 中的局部时间平均池层,  $\frac{k}{T}$  是时间特征间隔的一半。因此,相邻的 ltap 窗口重叠,以确保前后帧信息的相关性和完整性。

经过局部时间平均合并后,本步骤可以得到一系列单模态的短期特征。多模特征序列在输入 TCN

前融合成一个序列。整体短时时空特征模块在预训练时与截断后的流程如图 2 所示。

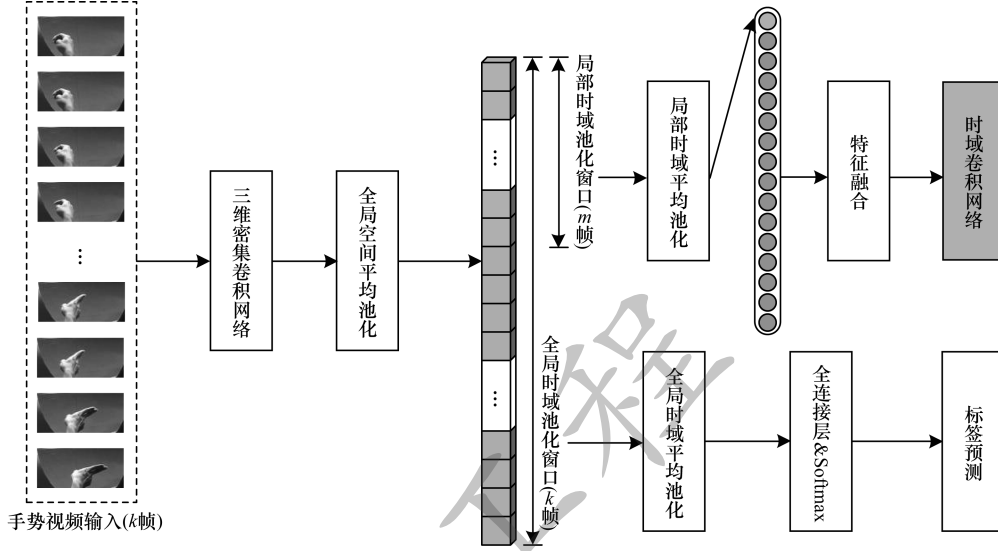


图 2 短时时空特征模块流程

Fig. 2 Procedure of short-term temporal and spatial feature module

## 2.2 短时时空特征序列的识别

基于从各种数据模式(RGB、光流、深度等)中提取的短时时空特征,考虑整个视频的长期时间特征,对给定手势进行分类。本文采用一种序列识别模型 TCNs,并对其进行了改进以处理长期时间信息。TCNs 的主要特点是使用因果卷积和将输入序列映射到相同长度的输出序列。此外,考虑到序列具有较长的历史,该模型使用了能够产生大的卷积野的膨胀卷积以及允许训练更深网络的残差连接。考虑到本文的任务是对手势视频类别进行分类,TCN 的输出层通过一个完全连接层进行进一步处理,得到每个手势序列的一个类标签。改进的 TCN 模型结构如图 3 所示。

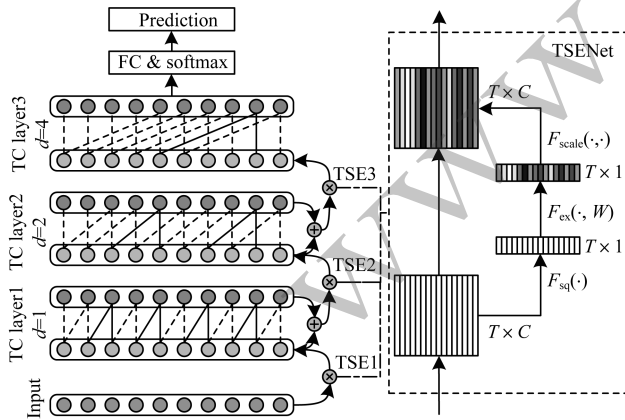


图 3 本文改进的 TCNtse 模型结构

Fig. 3 Structure of the improved TCNtse model proposed in this paper

在 TCN 模型中,从短时时空特征模块所学习得到的序列  $X = [x_1, x_2, \dots, x_t]$  作为 TCN 的输入序列,

其经过多层时间卷积后的输出设定为  $Y = [y_1, y_2, \dots, y_t]$ 。而每一层的卷积核本文都将使用膨胀卷积使得 TCNs 能够在不同层学习不同时序跨度的特征。膨胀卷积的计算公式为:

$$y_t = (x *_d h)_t = \sum x_{t-d_m} h_m \quad (4)$$

其中,  $*_d$  为膨胀卷积的运算符,  $d$  为卷积膨胀率,  $h$  为卷积核的参数。对于拥有  $L$  层的 TCN 模型而言,最后一层的输出  $y^L$  将被用来识别分类整个序列。整个序列的标签  $\hat{o}$  通过全连接、softmax 激活函数计算得到:

$$\hat{o} = \text{softmax}(W_o \times y^L + b_o) \quad (5)$$

其中,  $W_o$ 、 $b_o$  分别为训练后得到的全连接层参数。

值得注意的是,  $X = [x_1, x_2, \dots, x_t]$  中的特征在整个序列识别过程中的贡献是有所不同的。由于手势的组合特性和复杂性,在不同手势中必然会有一部分的手势片段是接近的,可将这些连续的手势片段进行识别区分,本文是通过 TCN 模型来学习手势片段间的时序关联性。但与此同时,如何学习到特征序列中的关联强度(权重)也是本文需要考虑的一个问题。为此,本文引入注意力机制,改进了压缩-激励网络(SENets)并将其应用嵌入到时序特征序列模型 TCN 中。

如图 3 所示,时域压缩-激励网络模块(TSENNet, TSE)被嵌入到 TCN 模块每一层时间卷积层输入前,首先将时间卷积层的输入  $X = [x_1, x_2, \dots, x_t]$  在通道上进行全局平均卷积,从而获得一条  $T \times 1$  大小的权值序列  $Z = [z_1, z_2, \dots, z_T]$ 。

假设时间卷积层输入通道数为  $C$ ,则  $t$  时刻的平均通道值  $z_t$  计算公式如下:

$$z_t = F_{sq}(x_t) = \frac{1}{C} \sum_{i=1}^C x_t(i) \quad (6)$$

与此同时,可以将平均通道所得到的值作为当前  $t$  时刻特征的权重,而本文为了重新调整各个时刻特征的权重,则加入了第 2 个操作,即压缩-激励操作。为使网络能够自动学习到这一权重值,加入了一个简单的激活门控制整个权值序列的计算:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \quad (7)$$

其中,  $\sigma$  为 sigmoid 激活函数,  $\delta$  为 ReLU 激活函数,  $W_1$  为大小是  $\frac{T}{r} \times T$  全连接参数,  $W_2$  为大小是  $T \times \frac{T}{r}$  全连接参数,而  $r$  为 TSENet 模块要压缩到的维度。将重调整权值序列  $s$  返回给原始输入  $X$  得到  $\tilde{X}$ :

$$\tilde{X} = F_{scale}(x_t, s_t) = s_t \cdot x_t \quad (8)$$

其中,  $\tilde{X} = [\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_T]$  即为经过权重再调配后的输入到时间卷积层的输入。

整个压缩-激励操作可以理解为一个编码-反编码的过程,将  $T \times 1$  的权值序列压缩至  $r \times 1$ ,而后再将其激励至  $T \times 1$  的序列,因为本文在每一步都加入了激励函数,所以在学习过程中,误差始终能够步步反传并调整参数  $W_1$ 、 $W_2$ ,从而获得重新调整权重后的权值序列  $s$ ,返回给原始输入  $X$  获得  $\tilde{X}$ ,并输入到 TCN 模型中,使得 TCN 模型在不改变原本结构的基础上,能够获得时序权值再调整后的短时时空特征序列进行学习并分类,提高准确度。

### 3 实验结果与分析

本文所提出的网络结构由 tensorflow 平台实现,并使用 NVIDIA Quadro gp100 GPU 进行训练。多模态截断的密集卷积网络 (T3D-Dense) 分别使用 RGB、深度信息和光流信息 (如果存在光流或可计算) 数据作为输入进行预训练。Adam 优化器用于训练 T3D-Dense 的预训练模型 3D-DenseNets,学习率初始化为  $6.4e-4$ ,每 25 个 epoch 下降 10 倍。权值衰减率设置为  $1e-4$ ,Drop\_out 设为 0.2,3D-DenseNets 内每个 block 的压缩率  $c$  和增长率  $k$  分别设置为 0.5 和 12。对于 TCN 模型,本文使用 Adam 优化器进行训练,学习率初始化为  $1e-4$ ,epsilon 为  $1e-8$ 。

#### 3.1 数据集

本节将本文方法与其他最新的动态手势方法进行比较。在实验中,本文使用两个公开的多模态动态手势数据集来评估文中提出的模型。

1) VIVA<sup>[15]</sup>。VIVA challenge 数据集是一个多模态的动态手势数据集,专门设计用于在真实驾驶环境中研究自然人类活动的复杂背景设置、不稳定照明和频繁遮挡等情况。此数据集是由微软 Kinect 设备捕获的,共有 885 个 RGB 和深度信息视频序

列,其中包括 8 名受试者在车内进行的 19 种不同的动态手势。

2) NVGesture<sup>[6]</sup>。NVGesture 数据集为了研究人机界面,采用多传感器多角度进行采集。它包含 1 532 个动态手势,这些手势是由 20 名受试者在一个有人工照明条件的汽车模拟器中记录下来的,这个数据集包括 25 类手势。它还包括动态 DS325 装置作为 RGB-D 传感器,用 DUO-3D 进行红外图像采集。在实验中,本文使用 RGB、深度和光流模态作为模型的数据输入,而光流图则使用文献[31]提出的方法从 RGB 流计算得到。

#### 3.2 数据预处理

数据预处理包括数据增强和数据规范化 2 个部分:

1) 数据增强。在 VIVA 数据集中,数据增强主要由 3 个增强操作组成,即基于视频帧的反序、水平镜像或同时应用前 2 个操作。如 VIVA 中有一类手势是在视频中从左往右移动,通过反序操作或者水平镜像,能够得到一个从右向左移动的手势作为从右向左移动的手势的增强。而同时应用反序和水平镜像操作,就能够得到一个从左往右移动的手势作为从左往右移动的手势的增强。

在 NVGesture 数据集中,每个视频图像被调整为  $256 \times 256$  像素的图像大小,然后用  $224 \times 224$  块随机裁剪,在裁剪时,同一数据的裁剪窗口在视频中位置不变。

2) 数据规范化。对于机器学习而言,数据规范化是必要的,特别是对于时序模型,时序上的量是固定的,所以对时序维度的重采样尤为重要。本文给定一个额定帧数  $k$ ,对于小于额定帧大小或大于额定帧数大小的视频,使用上采样和下采样的统一标准化来统一帧的数量。给定的视频  $V$  和  $n$  帧需要压缩或扩展到  $k$  帧,有以下 2 种情况:

(1) 当  $n > k$  时,将视频  $V$  平均分割为  $k$  节视频集  $V_s$ ,其中  $V_s = [V_1, V_2, \dots, V_k]$ 。对于视频集  $V_s$  中的每个片段,随机选择一个帧作为多个连续视频帧的表达。最后,将所有表示帧连接起来,并使它们成为规范化的结果。

(2) 当  $n < k$  时,选择视频中的  $n$  帧,然后通过前后复制插值的方式进行扩展,将原本  $n$  帧的视频片段扩展到  $k$  帧。

在 T3D-Dense 模型和 TCNs 模型中,输入数据的维数是固定的,具体来说,3D-DenseNet 在预训练过程中的所有输入的帧数都应该是固定的。经统计,VIVA 数据集的平均帧数  $k$  是 32 帧,NVGesture 数据集的平均帧数  $k$  是 64 帧,所以在实验中本文将 VIVA 数据集的  $k$  设置为 32,NVGesture 数据集的  $k$  设置为 64。

由于 C3D 计算的高复杂性,输入的视频图像像素大小被重采样为  $112 \times 112$  像素。

### 3.3 在 VIVA 上的测试结果

表 2 为在 VIVA 数据集的 RGB 和深度信息 2 个模态上测试的动态手势的性能。

表 2 本文方法与其他方法在 VIVA 数据集上正确率对比

Table 2 Accuracy of comparison of proposed method and other methods on the VIVA data set %

方法	融合模式	正确率
HOG + HOG2 <sup>[35]</sup>	RGB + Depth	64.50
CNN;LRN <sup>[15]</sup>	RGB + Depth	74.40
CNN;LRN;HRN <sup>[15]</sup>	RGB + Depth	77.50
C3D <sup>[14]</sup>	RGB + Depth	77.40
I3D <sup>[36]</sup>	RGB + Depth	83.10
MTUT <sup>[37]</sup>	RGB + Depth	86.08
3D-Dense(a)	RGB + Depth	88.21
Res3D + TCNs(b)	RGB + Depth	85.97
T3D-Dense + TCNs(c)	RGB + Depth	90.73
本文方法	RGB + Depth	91.54

实验结果表明,本文方法在 VIVA 数据集上获得了 91.54% 的正确率。从表 2 可以看出,本文提出的 T3D-Dense + TCNtse 在正确率上远优于 HOG + HOG2、CNN;LRN、CNN;LRN;HRN 以及 C3D 方法,分别高出 27.04%、17.14%、14.04% 以及 14.14%。而本文方法的识别正确率与 I3D 与 MTUT 方法正确率较为接近,这在一定程度上是由于 I3D 和 MTUT 与本文方法使用了较为相似的预训练方式。尽管如此,可以看到本文方法 RGB 和 depth 网络的性能在 I3D 和 MTUT 的基础上分别提高了 8.44% 和 5.46%。

同时,本文在 VIVA 数据集上测试了其他方法以证明各模块的有效性,测试的方法主要有:

1) 完整 3D-DenseNets。3D-DenseNets 预训练的过程其本质就是完整 3D-DenseNets 对动态手势的识别训练,所以可以直接对预训练的 3D-DenseNets 进行测试,测试识别正确率为 88.21%。

2) Res3D + TCNs。通过将短时时空特征提取模块的主干框架 T3D-Dense 改变为 Res3D 网络,可以发现基本的 T3D-Dense 作为主干框架在正确率上优于以 Res3D 为主干框架的 Res3D + TCNs 网络。并且本文 T3D-Dense + TCNs 网络的参数量仅为 141 万,而 Res3D + TCNs 网络的参数量为 4 535 万,是 T3D-Dense + TCNs 参数量的 30 倍之多,证明了本文算法的优越性。

3) T3D-Dense + TCNs 与本文方法之间的区别在于 TCN 网络中是否有 TSE 模块的嵌入,可以看到 TSE 模块的加入使得网络获得了 0.81% 的识别率提高。

本文统计了根据本文方法所得到的最终分类的混淆矩阵,如图 4 所示。

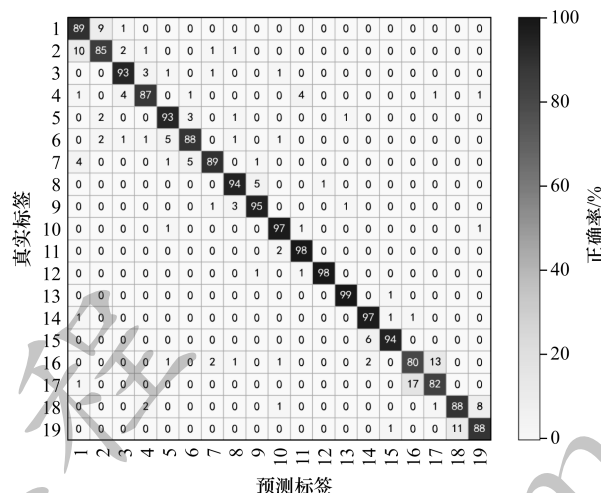


图 4 VIVA 数据集上输入为 RGB + 深度信息的识别混淆矩阵

Fig. 4 Confusion matrix with RGB + depth information input on VIVA data set

在实验中,发现在 VIVA 数据集上第 1 类与第 2 类、第 16 类与第 17 类上有着较高的误识别率,尤其是第 16 类与第 17 类(其中第 16 类为手势顺时针划圈,第 17 类为手势逆时针划圈)之间,误识别率为 15%。为此,从 TCN 各层中提取出第 16 类与第 17 类中 TSE 模块的权值作可视化。从图 5 中发现,由于第 16 类与第 17 类的结构空间信息在短时上拥有较多的相似性,导致 TSE 在权值控制上并不能很好地区分两者,使得两者在识别上会有较高的误识别率。但在大多数手势的权值上,尤其是 TCNs 第 3 层的 5 帧~12 帧上拥有较大的区分度。实验结果证明,TSE 对于 TCNs 识别具有较好的效果。

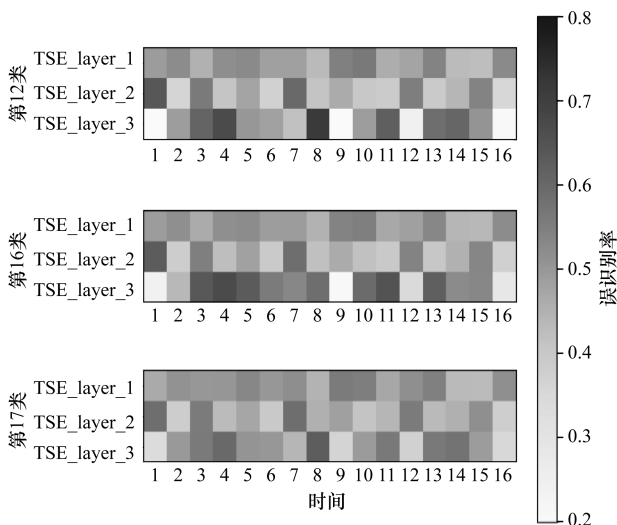


图 5 VIVA 数据集中 TSE 模块可视化图

Fig. 5 Visualization diagram of TSE module in VIVA data set



### 3.4 在 NVGesture 上的测试结果

为了在两种以上数据流的任务中测试本文的方法,在 NVGesture 数据集上分别以 RGB + 深度信息、RGB + 光流信息以及 RGB + 深度 + 光流信息作为输入进行了测试,分类结果如表 3 所示。

表 3 本文方法与其他方法在 NVGesture 数据集上正确率对比  
Table 3 Accuracy of comparison of proposed method and other methods on NVGesture data set %

方法	融合模式	正确率
HOG + HOG2 <sup>[35]</sup>	RGB + Depth	36.90
I3D <sup>[36]</sup>	RGB + Depth	83.82
MTUT <sup>[37]</sup>	RGB + Depth	86.10
本文方法	RGB + Depth	84.87
2S-CNNs <sup>[4]</sup>	RGB + Opt. flow	65.60
iDT <sup>[38]</sup>	RGB + Opt. flow	73.40
I3D <sup>[36]</sup>	RGB + Opt. flow	84.43
MTUT <sup>[37]</sup>	RGB + Opt. flow	85.48
本文方法	RGB + Opt. flow	86.21
R3DCNN <sup>[6]</sup>	RGB + Depth + Opt. flow	83.80
I3D <sup>[36]</sup>	RGB + Depth + Opt. flow	85.68
MTUT <sup>[37]</sup>	RGB + Depth + Opt. flow	86.93
本文方法	RGB + Depth + Opt. flow	86.37

在 RGB + 深度信息中,将本文方法与 HOG + HOG2、I3D 以及 MTUT 方法进行比较,可以看出,对于较为复杂的数据集,相较于传统 HOG + HOG2 方法,本文方法具有较高的正确率,但与 3D 与 MTUT 方法相比识别率不明显,甚至比 MTUT 正确率低 1.23%,可能是因为在复杂的数据集中,由于较为轻量的模型导致在时间跨度较大的数据中并不能很好地识别出视频片段间的联系。

在 RGB + 光流信息中,将本文方法与 2S-CNNs、iDT、I3D 以及 MTUT 方法进行比较。虽然 iDT 通常被认为是目前性能最好的手工识别方法,但可以看出,本文的方法识别正确率高于 iDT 方法 19.88%。并且在此模态中,由于光流信息同时包含了许多前后帧的手势变化信息,因此本文方法在精度上都高于其他方法。

在 RGB + 深度 + 光流信息中,将本文方法与 R3DCNN、I3D 以及 MTUT 方法进行比较。其中 R3DCNN 是该数据集原始方法,可以看出本文方法比原始方法正确率高 2.57%,比 I3D 方法高 0.69%。虽然本文方法正确率比最新的 MTUT 方法低 0.56%,但由于模型在特征融合上较为简单,因此结果在可接受范围内。在此基础上,本文方法在 NVGesture 数据集中各手势的误判率较为平均,平均误判率为 0.51%,识别混淆矩阵如图 6 所示。

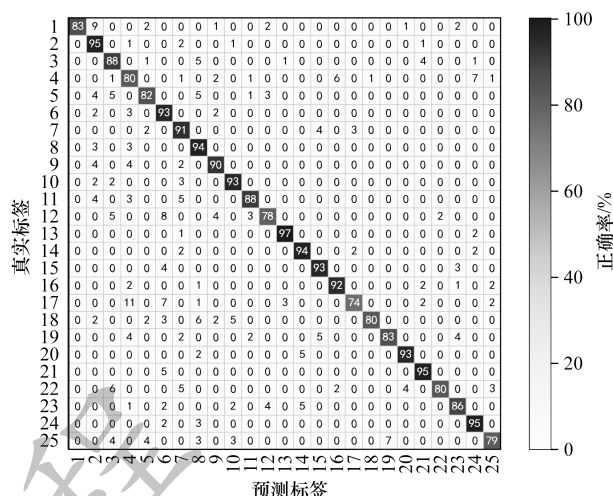


图 6 NVGesture 数据集输入为 RGB + 深度 + 光流信息的识别混淆矩阵

Fig. 6 Confusion matrix with RGB + Depth + Opt. flow information input on NVGesture data set

综上所述,本文方法在 NVGesture 数据集上的识别正确率取得了与当前最新方法相近的水平。

## 4 结束语

本文提出一种基于时域注意力机制的 Dense-TCN 模型。该模型通过截断预训练的 3D-DenseNets 和局部时域池化的方式来避免时间片段过多的重复训练,同时根据嵌入时域注意力机制改进 TSE 模块对短时时空特征序列进行识别。实验结果表明,该模型具有较高的识别率,且参数量较少。由于 3D-DenseNets 需要预训练且被截断才能提取局部的短时时空特征,依赖于预训练时 3D-DenseNets 的正确率和多模态融合方法,导致针对一些分类多、噪声大的数据集时正确率较低,因此下一步拟将 3D-DenseNets 模型改为端到端模型,并对多模态融合方法进行改进,以进一步提高模型识别率。

## 参考文献

- [1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 580-587.
- [2] RAUTARAY S S, AGRAWAL A. Vision based hand gesture recognition for human computer interaction: a survey[J]. Artificial Intelligence Review, 2015, 43(1): 1-54.
- [3] CAMGOZ N C, HADFIELD S, KOLLER O, et al. Subnets: end-to-end hand shape and continuous sign language recognition [C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 3075-3084.

- [4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. [2019-11-01]. <https://arxiv.org/abs/1406.2199>.
- [5] NEVEROVA N, WOLFC, TAYLOR G W, et al. Multi-scale deep learning for gesture detection and localization[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 474-490.
- [6] MOLCHANOV P, YANG X, GUPTA S, et al. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 4207-4215.
- [7] HUANG G, LIU Z, VAN DER M L, et al. Densely connected convolutional networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4700-4708.
- [8] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [EB/OL]. [2019-11-01]. <https://arxiv.org/abs/1803.01271v1>.
- [9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2019-11-01]. <https://arxiv.org/abs/1706.03762>.
- [10] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7132-7141.
- [11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[EB/OL]. [2019-11-01]. <https://blog.csdn.net/u011534057/article/details/51318670>.
- [12] WANG Pichao, LI Wanqing, LIU Song, et al. Large-scale isolated gesture recognition using convolutional neural networks [C]//Proceedings of the 23rd International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 7-12.
- [13] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015: 4489-4497.
- [14] MOLCHANOV P, GUPTA S, KIM K, et al. Hand gesture recognition with 3D convolutional neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1-7.
- [15] PIGOU L, VAN DEN O 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(2/3/4): 430-439.
- [16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2019-11-01]. <https://arXiv.org/abs/1409.1556>.
- [17] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1-9.
- [18] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [EB/OL]. [2019-11-01]. <https://arXiv preprint arXiv:1502.03167>.
- [19] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2818-2826.
- [20] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2017: 342-356.
- [21] HE Kaiming, ZHANG Xiaoyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.
- [22] LIN Chi, WAN Jun, LIANG Yanyan, et al. Large-scale isolated gesture recognition using a refined fused model based on masked Res-C3D network and skeleton LSTM [C]//Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Washington D. C., USA: IEEE Press, 2018: 231-246.
- [23] MIAO Q, LI Y, OUYANG W, et al. Multimodal gesture recognition based on the ResC3D network [C]//Proceedings of IEEE International Conference on Computer Vision Workshop. Washington D. C., USA: IEEE Press, 2017: 675-689.
- [24] ESCALANTE H J, VICTOR P L, WAN J, et al. ChaLearn joint contest on multimedia challenges beyond visual analysis: an overview [C]//Proceedings of the 23rd International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 453-468.
- [25] WAN Jun, LIN Chi, XIE Yiliang, et al. Results and analysis of ChaLearn LAP multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges [C]//Proceedings of IEEE International Conference on Computer Vision Workshop. [S. l.]: IEEE Computer Society, 2017: 469-478.
- [26] ZHANG Tong, WANG Rong, DING Jianwei, et al. Face recognition based on densely connected convolutional networks [C]//Proceedings of the 4th IEEE International Conference on Multimedia Big Data. Washington D. C., USA: IEEE Press, 2018: 1-6.
- [27] ZHANG Jian, ZHANG Yonghui, HE Jingxuan. Action recognition algorithm based on DenseNet and depth motion map[J]. Information Technology and Network Security, 2020, 39(1): 63-69. (in Chinese)  
张健, 张永辉, 何京璇. 基于 DenseNet 和深度运动图的行为识别算法[J]. 信息技术与网络安全, 2020, 39(1): 63-69.
- [28] CAO Chuqing, LI Ruifeng, ZHAO Lijun. Hand posture recognition method based on depth image technology[J]. Computer Engineering, 2012, 38(8): 16-18, 21. (in Chinese)  
曹雏清, 李瑞峰, 赵立军. 基于深度图像技术的手势识别方法[J]. 计算机工程, 2012, 38(8): 16-18, 21.



- [29] YI Sheng, LIANG Huagang, RU Feng. Hand gesture recognition based on multi-column deep 3D convolutional neural network[J]. Computer Engineering, 2017, 43(8): 243-248. (in Chinese)  
易生,梁华刚,茹锋. 基于多列深度 3D 卷积神经网络的手势识别[J]. 计算机工程, 2017, 43(8): 243-248.
- [30] ZHANG Liang, ZHU Guangming, SHEN Peiyi, et al. Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition [C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 3120-3128.
- [31] CHAI Xiujuan, LIU Zhipeng, YIN Fang, et al. Two streams recurrent neural networks for large-scale continuous gesture recognition [C]// Proceedings of the 23rd International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 31-36.
- [32] HOU Jiangxun, WANG Guilin, CHEN Xinghao, et al. Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition [C]// Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 562-579.
- [33] DING Chongyang, LIU Kai, LI Guang, et al. Spatio-temporal weighted posture motion features for human skeleton action recognition research [J]. Chinese Journal of Computers, 2019, 43(1): 29-40. (in Chinese)  
丁重阳,刘凯,李光,等. 基于时空权重姿态运动特征的人体骨架行为识别研究 [J]. 计算机学报, 2019, 43(1): 29-40.
- [34] XIE Zhao, ZHOU Yi, WU Kewei, et al. Activity recognition based on spatial-temporal attention LSTM [EB/OL]. [2019-11-01]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20191227.1658.002.html>. (in Chinese)  
谢昭,周义,吴克伟,等. 基于时空关注度 LSTM 的行为识别 [EB/OL]. [2019-11-01]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20191227.1658.002.html>.
- [35] OHN-BAR E, TRIVEDI M M. Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations [J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(6): 2368-2377.
- [36] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 6299-6308.
- [37] ABAVISANI M, JOZE H R V, PATEL V M. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 1165-1174.
- [38] WANG H, ONEATA D, VERBEEK J, et al. A robust and efficient video representation for action recognition [J]. International Journal of Computer Vision, 2016, 119(3): 219-238.

编辑 索书志