**文章编号**:1671-4598(2020)04-0196-04　　DOI:10.16526/j.cnki.11-4762/tp.2020.04.040　　**中图分类号**:TP391　　**文献标识码**:A

# 基于语义分割与迁移学习的手势识别

邢予权，潘今一，王　伟，刘建烽

(浙江工业大学 信息工程学院，杭州　310023)

**摘要**：针对复杂场景下深度相机环境要求高，可穿戴设备不自然，基于深度学习模型数据集样本少导致识别能力、鲁棒性欠佳的问题，提出了一种基于语义分割的深度学习模型进行手势分割结合迁移学习的神经网络识别的手势识别方法；通过对采集到的图像数据集进行不同角度旋转、翻转等操作进行数据集样本增强，训练分割模型进行手势区域的分割，通过迁移学习卷积神经网络更好地提取手势特征向量，通过 Softmax 函数进行手势分类识别；通过 4 个人在不同背景下做的 10 个手势，实验结果表明：针对复杂背景环境下能够正确地识别手势。

**关键词**：语义分割；迁移学习；手势识别；卷积神经网络

# Gesture Recognition Based on Semantic Segmentation and Transfer Learning

Xing Yuquan, Pan Jinyi, Wang Wei, Liu Jianfeng

(College of Information Engineering, Zhejiang University of Technology, Hangzhou　310023, China)

**Abstract**: Due to the high requirements for the deep camera environment in complex scenes, wearable devices are not natural, and the lack of data set samples based on the deep learning model leads to poor recognition ability and robustness, A gesture recognition method based on deep learning model based on semantic segmentation and neural network based on transfer learning is proposed. By rotating and flipping the collected image data set at different angles, data set samples were enhanced, segmentation model was trained to segment gesture areas, and gesture feature vectors were extracted better through transfer learning convolutional neural network. Softmax function is used for gesture classification and recognition. Through 10 gestures made by 4 people in different backgrounds, the experimental results show that they can correctly recognize gestures in complex environments.

**Keywords**: semantic segmentation; transfer learning; gesture recognition; convolutional neural networks

## 0　引言

人与人之间的沟通其实就是两者之间互相交换信息的过程，我们在平时生活当中有很多种交流的方式，比如：我们可以用我的微笑、眨眼、手势等方式将想要表达的信息传达给外界，同时也可以通过我们的眼睛、触觉等发现四周别人给我们传递的消息，例如：别人打手势、肢体之间接触等。所谓人人交互，顾名思义就是通过这些各种不同的方式进行人与人之间交换信息，同理，人机交互（human－computer interaction，HCI）就是人和计算机之间通过某种信息交换方式（例如：手势）实现对一些机器之间的控制的目的，当前人机交互主要研究的内容是如何进行人和计算机进行交换信息的方式，传统的交互方式主要是采用专用设备进行，例如：键盘、鼠标、触摸屏等输入设备，需要提前制定一些信息交换规则，按照规则进行与计算机之间互换信息，导致计算机处于被动接受信息的状态。随着当下技术的发展与不断进步，人们已经开始改变传统被动接收信息，向着主动方式发展，随着计算机视觉技术的越来越成熟，人们的生活方式也逐渐发生了变化，对生活的品质要求也发生改变，越来越火的人机交互受到人们的热捧。手势作为一种不用说话就可以向他人传达一些信息的肢体语言，在生活当中很多环境下给予了大量帮助，手可以通过与身体其他部位的组合可以产生很多的表达信息，比如：聋哑残障人士的交流就是通过手语进行交互，同时手势也可以和机器之间交流等等。由于手势是非常的灵活，不同的时间或者地点可能表达的意思就完全不同。例如，美国和欧洲可能认为对于别人指向伸出的手指是很正常的交流，但对于亚洲人来说他被定义为一种不礼貌的行为。静态手势从理论上来说指可以认为手的个体形状，它不考虑在具体的哪个时间和方向，位置信息，只包括单个手的形状，例如卡住拇指和小指，伸直其他三个指头，就表示数字三。因此，手势识别也成为人机交互技术研究领域的主流方向之一[1]。

传统的方法是基于穿戴设备[2-4]（例如：数据手套）获取手部的位置信息和旋转信息作为手势信息，虽然能够在复杂的背景和光照[5]下获取较好的准确性和稳定性，但是设备的成本太高，人的手势约束较多，在人机交互过程中

**收稿日期**:2019-09-11；　**修回日期**:2019-10-15。

**作者简介**:邢予权(1991-),男,河南巩义人,硕士研究生,主要从事图像处理、模式识别方向的研究。

潘今一(1959-),男,浙江湖州人,博士,教授,主要从事模式识别、图像处理方向的研究。

不能够自然地进行交互。针对自然性交互与成本问题，基于相机采集越来越受到青睐，基于深度相机[6-7]能够获取人体部位多维信息（例如：RGB，深度信息），环境条件要求高，采集范围小，采用普通相机采集，张彩珍[8]等人通过肤色模型与 BP 神经网络手势识别的方法，杨红玲[9]等人提出一种基于 YCbCr 空间肤色分割去除背景结合卷积神经网络进行手势识别，其利用人体肤色与周围环境差异快速实现手势的分割，易受到光照和复杂背景的影响，类肤色或者人脸等区域的噪声干扰，从而使基于肤色分割的阈值很难设置，从而导致手势分割较差，卷积神经网络识别数据样本少，识别率低。针对上述问题，提出采用语义分割深度学习模型[10-12,14]手势分割，降低基于肤色模型等方法面对复杂场景下出现的难以获取手势区域问题，通过迁移学习[14]的方法来处理工程实践中遇到样本数据少等问题，从而增强图像识别的准确率。

## 1　基于语义分割进行手势分割

因为佩戴设备或者采用深度相机的很多局限性原因，当前采用普通相机进行手势识别研究受到广大研究者的青睐。采用普通相机的静态手势识别系统，主要是通过手的形状的分离与形状特征提取分类两大模块，检测的主要方式是通过一些算法（差分法、肤色分割等）将手型的区域凸出来，而非手势的区域遮挡掉，这样手势区域非手势区域有一个较大的凸出对比，这样对于后面的特征提取有很大的影响，因此将手势分割作为手势识别的第一步是非常重要的，它的好坏对后面的手势识别的识别准确率有很大影响[8-10]。传统的识别方法利用直方图的分割、局部区域信息的分割、颜色空间等特征进行分割，然后，手动提取一些特征采用 SVM 等分类器识别，但是传统分割和特征提取技术往往会受到复杂光照、背景等一些影响。随着深度学习的发展，应用于图像的分割逐渐成为一种主流的图像识别的预处理方法，在深度学习领域基于像素处理的语义分割近些年得到快速发展，大多数都是全卷积神经网络[11,14]（Fully convolutional networks，FCN）为基础不断的更新升级换代产生的一系列基于卷积神经网络的语义分割方法。

根据分割算法在公共数据集的表现，采用 Encoder－Decoder with Atrous Separable Convolution for Semantic Image Segmentation（deeplabv3＋）[11]算法进行手势分割。

### 1.1　deeplabv3＋算法

语义分割 Deeplab v3＋网络采用“编码－解码”结构。编码模块使用改进的 Xception 网络进行特征提取，其后连接 ASPP 模块用于提取并融合图像的多尺度特征。解码模块将编码模块特征图与改进的 Xception 中间特征图进行融合，然后上采样得到分割结果[15-16]。

Deeplabv3＋网络架构分别如图 1 所示。图像首先经过改进的 Xception 网络得到宽高为原图 1/1 大小的特征图，接着进入孔洞空间金字塔池化 ASPP 模块，使用 4 个并行的具有不同孔洞率的孔洞卷积对特征图进行特征提取，获得 4 个具有不同感受视野的特征图。使用全局平均池化对特征图进行融合，使得编码模块网络最后的特征图融合了图像的多尺度信息，有助于提高小目标物体的分割结果。编码模块输出的特征图虽然能够编码丰富的语义信息，但多次下采样步骤和不同步长的卷积计算会导致物体边界信息的丢失，直接上采样到原始图像尺寸的语义分割图片只能得到粗略的物体边界，精度较低。因此，解码模块先将编码模块最后一层特征图上采样 4 倍，与编码模块改进的 Xception 网络中下采样 1/4 的特征图进行融合，然后再次 4 倍上采样得到最终包含精确物体边界的语义分割图片。

受限于硬件环境能力，实际训练中的不训练时间可能会很长，并且训练样本规模数量有限，如果随机给定初始化模型参数权值，训练效果不一定良好，使用 Xception 的骨架网络通过大量公开数据集进行的预训练好的模型，作为进行网络训练时的初始化的参数。通过迁移学习可加速网络训练速度，提高训练精度，减少训练时间。

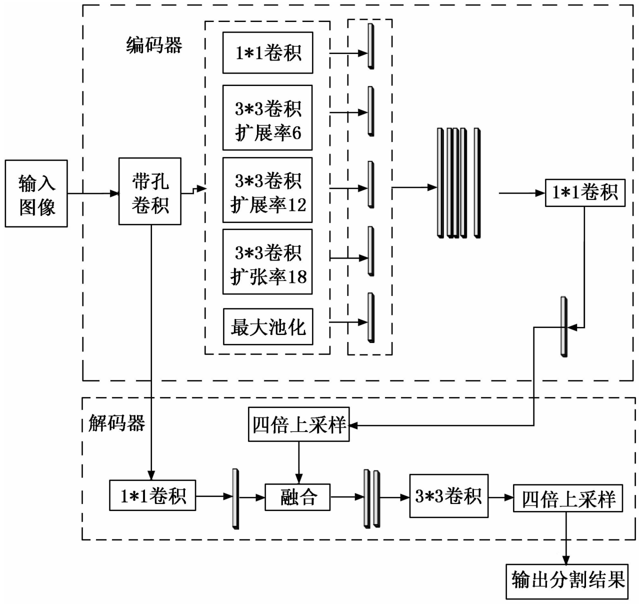


图 1　deeplabv3＋网络结构

### 1.2　空洞卷积

传统的卷积神经网络对图像操作时首先对输入图像进行卷积再进行池化操作，降低输入图像的尺寸大小同时增大感受视野，在池化降低尺寸增大感受视野，向上采用恢复原图大小过程中出现像素的丢失，因此采用空洞卷积同时可以同时进行两步操作。如式（1）所示：

$$y[i] = \sum_{k} x[i + r \cdot k] w[k] \tag{1}$$

式（1）中，$i$ 为图像像素位置，$w$ 为卷积核的大小，速率 $r$ 是采样点之间加入 $r-1$ 个零。

## 2　手势识别

自卷积神经网路提出以来，人们在不断研究过程当中

发现其对图像具有很强的特征提取性能，被广泛应用在图像识别、图像分类等各方面，手势识别采用该网络主要通过多层卷积与采用来提取特征，通过分类函数将特征进行分类识别，常用的卷积神经网络主要有以下几部分组成，如图 2 所示。

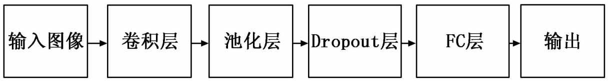


图 2 卷积网络基本结构

### 2.1 卷积层

卷积层就是利用卷积核与图像局部按照一定关系进行卷积操作提取局部视野特征，作为下一层的输入继续操作，逐渐获取图像由低到高的层级特征。假设图像输入大小为（$X*Y$），卷积公式如式（2）所示：

$$x_{ab}=f\left(\sum_{j=0}^{w-1}\sum_{i=0}^{h-1}x_{a+i,b+j}w_{ij}+C\right) \tag{2}$$

式中，卷积核为 $w*h$，$x_{a+i,b+j}$是输入图像内像素坐标，$w_{ij}$是卷积核（$i$，$j$）处的值，$C$ 为偏置，$x_{ab}$就是卷积结果。

### 2.2 池化层

池化层又名下采样。提取局部视野的主要特征，丢弃次要特征，降低了特征的维数同时减少模型参数的数量，从而使主要信息得以保留。

### 2.3 Dropout 层

在训练模型过程当中，为了防止出现过拟合现象，根据生物学原理，每次进行神经元激活时并不是所有都唤醒，因此在训练过程当中每次随机的丢弃一些节点，这样每次训练的网络不同，防止出现拟合来提高神经网络的性能。

### 2.4 全连接层

最后要输出待分类目标出现的概率，采用 softmax 分类函数进行目标分类。计算目标在各类中出现的概率计算如式（3）所示：

$$p_i=\frac{exp(X_i^L)}{\sum_{j=1}^{k}\exp(X_j^L)} \tag{3}$$

式中，假设隐藏层 $L$ 输出是 $K$ 维矩 $p_i$ 是出现在各类别概率，$X_i^L$ 是 $L$ 层得到的 $K$ 维矩阵第 $i$ 个分量。

采用交叉熵损失函数来计算预测值与实际值差异。如式（4）：

$$L=-\frac{1}{n}\sum_{n-1}^{n}\log P(y^i\mid x^i;\omega) \tag{4}$$

式中，$n$ 为待训练目标，$y^i$ 是各个对象的实际标签类型。

## 3 迁移学习

为了训练过后能在目标数据集上取得良好的分类效果，具有足够的样本数量是一个稳定的保障，但是实际生活当中获取合适的样本数据非常困难，往往需要大量的人力与物力，同时传统的机器学习方法要求训练集和测试集具有相同的特征空间和数据分布，因此导致数据样本的充足更是难上加难，然而过去耗费大量费用制作的相似数据集又不能使用，导致资源的浪费，因此采用迁移学习的目的就是利用以前过时数据样本作为预训练模型，来增加网络训练过程当中网络的稳定性，提高网络特征提取能力，然后利用预训练提取特征的某部分参数，采用自己的数据集训练高层特征提取与分类层参数，从而可以更好地解决训练样本缺失所导致的局部最优解和过拟合等问题。本算法主要利用公开数据集 ImageNet 具有大量的样本能够更好地训练网络提取特征的能力，本文采用 VGG16 网络在 ImageNet 数据集上训练的预训练模型进行网络初始化进行提取特征本数据单独进行训练分类层网络模型。VGG16 网络如表 1 所示。

表 1 VGG16 网络

| Layer(type) | Output shape | Param # |
|---|---|---|
| Input_1(InputLayer) | (None,480,640,3) | 0 |
| Block1_conv1(Conv2D) | (None,480,640,64) | 1792 |
| Block1_conv2(Conv2D) | (None,480,640,64) | 36928 |
| Block1_pool(MaxPooling2D) | (None,240,320,64) | 0 |
| Block2_conv1(Conv2D) | (None,240,320,128) | 73856 |
| Block2_conv2(Conv2D) | (None,240,320,128) | 147584 |
| Block2_pool(MaxPooling2D) | (None,120,160,128) | 0 |
| Block3_conv1(Conv2D) | (None,120,160,256) | 295168 |
| Block3_conv2(Conv2D) | (None,120,160,256) | 590080 |
| Block3_conv3(Conv2D) | (None,120,160,256) | 590080 |
| Block3_pool(MaxPooling2D) | (None,60,80,256) | 0 |
| Block4_conv1(Conv2D) | (None,60,80,512) | 1180160 |
| Block4_conv2(Conv2D) | (None,60,80,512) | 2359808 |
| Block4_conv3(Conv2D) | (None,60,80,512) | 2359808 |
| Block4_pool(MaxPooling2D) | (None,30,40,512) | 0 |
| Block5_conv1(Conv2D) | (None,30,40,512) | 2359808 |
| Block5_conv2(Conv2D) | (None,30,40,512) | 2359808 |
| Block5_conv3(Conv2D) | (None,30,40,512) | 2359808 |
| Block5_pool(MaxPooling2D) | (None,15,20,512) | 0 |

## 4 实验结果与分析

实验测试环境：采用戴尔笔记本电脑 Intel（R）Core（TM）i5－7300 2.50 GHz CPU 处理器，采用 Visual Studio 2015，OpenCV3.2.0，tensorflow1.12，python3.5，显卡 GTX1050ti 实现了手势识别的方法。

### 4.1 语义分割实验及结果

1）采用 labelme 开源标注工具，制作满足图像分割要求数据集标签如图 3 所示。

图 3 标记标签制作

2）deeplabv3＋ 与基于肤色模型分割对比如图 4 所示。

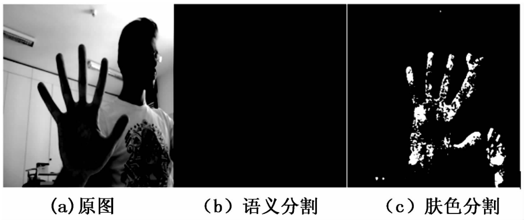


图 4　分割对比图

### 4.2　手势识别的结果分析

本文使用在 ImageNet 上预训练得到的网络模型对 VGG16 模型进行特征提取层初始化操作，提取我们所做的数据集特征进行分类模型训练。我们将初始学习率设为 0.001，保存迭代 100 次（全部数据集训练一遍为一次迭代）将训练结果进行实验对比分析。

1）本次采用的数据集是 Senz3D 公开数据集进行手势识别，由于数据集场景复杂性较好，但是样本相对减少采用样本增强方式有如下：

（1）对图像进行 30°、－30°、－60°、60°、45°、－45°等操作对样本数据集进行扩充。

（2）采用风格迁移进行其他背景迁移到手势图像作为背景进行图像特征的多样化。

（3）通过模拟噪声的分布，对图像增加各种噪声，增加扰动信息，达到样本复杂多样性，以便更好地提高泛化能力。

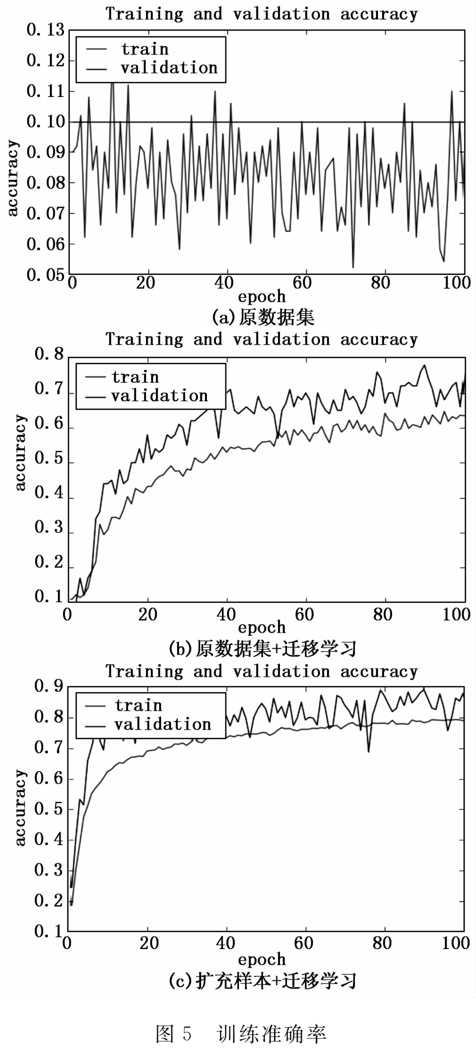


图 5　训练准确率

2）准确率曲线分析：分别进行原数据集进行训练，基于迁移学习对原数据集进行训练，扩充样本后迁移学习进行实验，如图 5 所示。

通过观察以上三幅曲线图，图 5（a）明显出现没有收敛趋势，这样的训练结果没有学习的意义，图 5（b）经过迁移学习之后出现逐渐收敛的趋势，准确度明显提高，说明迁移学习对于小样本学习具有较好的表现，通过图 5（c）采用迁移学习与样本增强的方法后，准确度又有了一定的提高，实验结果表明，面对小样本数据时通过迁移学习与样本扩充方法有一定的提高。

## 5　结束语

针对小样本复杂背景下的手势识别问题，采用普通相机有效解决了图像采集中可穿戴设备传感器费用高、深度相机环境要求高缺点，通过与传统的不同颜色分割模型分割能力进行对比，深度学习的图像分割技术可以替代传统肤色分割模型等技术分割，避免出现的阈值难找，受到光照、类肤色等环境影响，导致分割中出现类肤色出现或者较少非手势肤色区域出现；减少很多分割图的预处理运算，具有很好的普适性，同时采用迁移学习的方法与扩充样本相结合准确度表现来看可以更好地防止样本数量少时出现的过拟合现象，可以减少数据的收集处理成本，可以有很好的工程使用价值，同时也面临一些不足之处，有的场景下旋转角度并不适合扩充样本，没有很好的普适性，需要朝着增加数据样本多样性算法进一步研究，增加普适性。

**参考文献：**

[1] 易靖国，程江华，库锡树. 视觉手势识别综述 [J]. 计算机科学，2016，43 (z1)：103－108.

[2] 都明宇，王志恒，等. 基于多通道 sEMG 小波包分解特征的人手动作模式识别方法 [J]. 计算机测量与控制，2018，26 (6)：160－163.

[3] 王景芳，施　霖. 基于人工神经网络对 sEMG 信号的手势动作识别 [J]. 传感器与微系统，2017，36 (6) ：63－65.

[4] 张发辉，杨大勇，等. 基于肌电信号和姿态信号的手势识别 [J]. 传感器与微系统，2019，38 (7)：46－49.

[5] 邓志敏. 基于复杂背景下的手势识别系统 [D]. 桂林：广西师范大学，2018.

[6] 王　兵，董洪伟，张明敏，等. 基于 Kinect 的动态手势识别 [J]. 传感器与微系统，2018，37 (2)：143－146.

[7] 谭台哲，韩亚伟，邵　阳. 基于 RGB－D 图像的手势识别方法 [J]. 计算机工程与设计，2018 (2)：511－515.

[8] 张彩珍，张云霞，等. 基于肤色模型与 BP 神经网络的手势识别 [J]. 传感器与微系统，2019，38 (6)：140－143.

（下转第 204 页）