计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(6):47−55 [doi: 10.15888/j.cnki.csa.007448]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于肤色分割与改进 VGG 网络的手语识别①

包嘉欣, 田秋红, 杨慧敏, 陈影柔

(浙江理工大学 信息学院, 杭州 310018)
通讯作者: 包嘉欣, E-mail: bbaojiaxin@163.com

摘　要: 传统的手语识别仅仅依靠人工选取的底层特征完成识别, 难以适应手语图像背景的多样性, 本文提出了一种综合多要素的手语肤色分割与改进 VGG 网络的手语识别方法. 对采集到的手语图像利用椭圆模型进行初步分割, 根据最大连通域排除背景中的类肤色区域并用质心定位的方法去除手部区域以外的肤色区域, 从而实现手语图像准确分割. 在原有 VGG 网络的基础上减少卷积及全连接的层数对 VGG 网络进行改进, 减少了所需的存储容量和参数数量. 将分割后的手语灰度图像作为网络的输入, 采用改进的 VGG 网络建立手语的识别模型. 通过比较不同结构的网络模型对手语图像的识别率, 表明改进的 VGG 网络能够有效进行特征学习, 对手语图像的平均识别率都达到 97% 以上.

关键词: 肤色分割; 手语识别; VGG; 改进 VGG 网络; 识别模型

引用格式: 包嘉欣,田秋红,杨慧敏,陈影柔.基于肤色分割与改进 VGG 网络的手语识别.计算机系统应用,2020,29(6):47–55. http://www.c-s-a.org.cn/1003-3254/7448.html

Sign Language Recognition Based on Skin Color Model and Improved VGG Network

BAO Jia-Xin, TIAN Qiu-Hong, YANG Hui-Min, CHEN Ying-Rou

(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The traditional sign language recognition only relies on the underlying features selected manually, which is difficult to adapt to the diversity of sign language image background. A method of sign language recognition based on the multi-factor skin color segmentation and the improved VGG network is proposed in the study. The collected sign language images are initially segmented by an elliptic model. The skin color region is excluded according to the maximum connected domain, and the skin color regions outside the hand region is removed by centroid positioning method, so as to realize the accurate segmentation of sign language images. The VGG network is improved by reducing the number of convolution and full connection, which reduces the required storage capacity and the number of parameters. The gray-scale image of the segmented sign language is taken as the input of the network, and the improved VGG network is used to establish the recognition model of sign language. By comparing the different structure of the network model of sign language recognition rate of the image, show that the improved VGG networks can effectively study characteristics, the average image sign language recognition rate is above 97%.

Key words: skin color segmentation; sign language recognition; VGG; improved VGG network; recognition model

① 基金项目: 国家自然科学基金 (51405448); 浙江理工大学博士科研启动项目 (11122932611817); 浙江省大学生科技成果推广项目 (14530031661961); 浙江理工大学 2019 年国家级大学生创新创业训练计划 (201910338012)
Foundation item: National Natural Science Foundation of China (51405448); Scientific Research Start-Up Fund for Doctorate of Zhejiang Sci-Tech University (11122932611817); Graduates Scientific Research Achievement Promotion Program of Zhejiang Province (14530031661961); Year 2019, Innovation Training Program for Students of Zhejiang Sci-Tech University (201910338012)
收稿时间: 2019-11-09; 修改时间: 2019-12-09; 采用时间: 2019-12-17; csa 在线出版时间: 2020-06-10

© 中国科学院软件研究所 http://www.c-s-a.org.cn

手语的构成主要是借助手和手臂完成的手势语, 是包含信息量最多的一种人体语言, 它与口语及书面语等自然语言的表达能力相当. 手语识别技术提供一种更为简单自然的人机交互方式, 它逐渐改变着人们的生活方式, 并已广泛应用于体感游戏、机器人控制、智能家电和车载系统等领域, 其研究发展影响着人机交互的自然性和灵活性, 具有重要的社会经济价值和研究意义. 手语识别不仅是听力障碍者的主要交流手段, 而且有效的手语识别将减轻听力障碍者因交流不便带来的困扰, 因此手语识别具有重要的社会意义.

根据手语识别提取特征的方法不同, 手语识别主要分为以下几类: 1) 基于穿戴式输入设备的识别方法[1–4], 该方法利用穿戴式的设备采集手的位置、形状和运动轨迹和运动方向等信息, 获得的手势时序可直接用于分类器识别. 但是该方法要求穿戴的设备比较昂贵, 且易损坏, 不容易维护, 难以推广和普及. 2) 基于人工设计特征的识别方法[5–8], 该方法利用通过提取合适的手语特征作为识别特征, 但是该方法的学习能力不强, 在样本量不断增大的情况下, 识别率不会显著提高, 且提取的特征容易受到光照、背景的影响. 3) 基于神经网络的识别方法[9–12], 该方法基于统计的方法能够实现复杂的非线性映射, 且具有分类特性和抗干扰性, 但是该方法在手语图像不足的情况下, 容易陷入过拟合.

基于深度学习的卷积神经网络具有结构层次化、权值共享、区域局部感知、特征提取和识别分类相结合的全局分类特点, 能够逐层自动地学习到合适的特征并进行分类, 在图像识别领域获得了广泛的应用. Liu 等[13]提出了基于深度神经网络的转移学习算法来解决带标记的彩色图像样本不足的问题, 与原始的 VGG 方法和浅层机器学习方法相比, 提出的方法具有更高的精度. Gu 等[14]提出将复杂算法 (卷积和批量归一化) 应用于 VGG 网络, 并对模型进行了扩展, 通过训练具有相同网络结构的实值 VGG 网络和复值 VGG 网络, 得到了训练和测试的精度. Ha 等[15]提出了一种基于图像的建筑信息模型 (BIM) 和 VGG 的室内定位新方法. 该方法通过渲染 BIM 图像构建数据集, 并在数据集中搜索与室内照片最相似的图像, 从而估算出照片的室内位置和方向, 结果证明了 VGG 网络中的池化层适合于特征选择. 但是 VGG 网络模型对手语图像数据集的数量要求过高, 且在训练模型时需要大量的存储容量, 对硬件的要求较高.

针对以上不足, 本文提出了一种基于肤色分割与改进 VGG 网络的手语识别方法. 在保证识别准确率的同时解决了复杂背景下手势图像的特征提取问题. 通过优化网络结构, 减少了模型所需的存储容量和参数量.

1 综合多要素的手语肤色分割方法

目前, 基于视觉的手势分割算法主要有基于肤色的手势分割算法、基于轮廓的手势分割算法和基于运动的手势分割算法. 基于肤色分割的方法, 通过在原始图像中选取与手部皮肤颜色相近的像素点, 然后把这些像素点所在的区域分割出来. 基于肤色分割的方法简单高效, 不受尺度和角度等因素的影响, 得到了广泛的应用. 但是基于肤色分割的方法容易受到背景中类肤色区域的干扰, 本文提出了一种综合多要素的手语肤色分割方法. 该方法首先采用椭圆模型对手语图像进行初步分割, 然后利用基于最大连通域和质心定位的方法来排除背景中的类肤色区域及除手部区域以外的肤色区域, 进而分割出手部区域.

1.1 基于椭圆模型的肤色分割

由于肤色对人的表情、动作等变化具有强烈的抗干扰能力, 因此常常将它作为手语识别与人脸识别的首选特征, 不同光照变化会导致肤色的亮度发生变化, 需要选择一个可靠的肤色模型来检测肤色区域[16]. YCbCr 颜色空间的是一种能将亮度信号和色度信号单独分离开的颜色空间, 其中 Y、Cb、Cr 分别指亮度、蓝色色度、红色色度. 从 RGB 颜色空间到 YCbCr 颜色空间的转换公式如下[17]:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{1}$$

其中, R、G、B 值分别为图像红、绿、蓝颜色值归一化后的值.

采集手语图像数据集中肤色的样本点, 并将肤色转化到 YCbCr 颜色空间, 然后在 CbCr 平面进行投影, 得到一个 CbCr 的椭圆, 判断坐标(Cb,Cr)是否在椭圆内 (包括边界), 即可判断是否为肤色像素点, 进而形成的统计椭圆模型如下:

$$\frac{(x-eC_x)^2}{a^2} + \frac{\left(y-eC_y\right)^2}{b^2} = 1 \tag{2}$$

© 中国科学院软件研究所 http://www.c-s-a.org.cn

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} C_b' - C_x \\ C_r' - C_y \end{bmatrix} \tag{3}$$

其中, C_x=109.38, C_y=152.02, a=25.39, b=14.03, θ=2.53, eC_x=1.60, eC_y=2.41.

1.2 图像去噪

经过肤色分割后, 手语图像中可能会存在孤立的噪声点和小的干扰块 (类肤色背景), 且肤色区域会存在大小不一的孔洞, 这些因素会严重干扰手部区域的提取, 因此必须去除.

中值滤波法是一种非线性平滑技术, 它将每一像素点的灰度值设置为该点某邻域内的所有像素点灰度值的中值, 对毛刺和孔洞的填充具有重要作用[18]. 在图像滤波中最常用3×3的窗口对图像进行中值滤波, 即选取指定点周围的 8 邻域的像素值进行排序, 将排序后的中值作为指定点的像素值. 中值滤波的公式如下:

$$g(i,j) = med\begin{bmatrix} f(i-1,j-1) & f(i-1,j) & f(i-1,j+1) \\ f(i,j-1) & f(i,j) & f(i,j+1) \\ f(i+1,j-1) & f(i+1,j) & f(i+1,j+1) \end{bmatrix} \tag{4}$$

其中, $f(i,j)$为原图像的像素值, $g(i,j)$为中值滤波后像素值, med为中值运算符.

漫水填充算法是一种用颜色来填充连通区域的算法, 首先从连通域里选出一点, 将该点作为种子点, 然后从该点开始寻找当前的连通域内其他的点, 并将这些点填充成指定的颜色.

本文先采用中值滤波对肤色分割后的手语图像进行平滑滤波, 去除孤立的噪声点及边缘的毛刺, 然后采用漫水填充算法填充肤色区域的孔洞, 确保手语区域的完整性.

1.3 基于最大连通域和质心定位的手部区域获取

经过肤色分割和图像去噪后, 图像中仍存在 3 处皮肤区域及其他稍微大一点的类肤色背景区域. 本文提出了一种基于面积算子和质心位置的手部区域定位方法, 实现了手部区域的获取.

计算图像中每个连通区域内的像素数目, 找出最大的 3 个连通区域, 即脖子区域、手臂区域、和手部区域, 舍弃其他连通区域.

根据式 (5)~式 (7) 计算 3 个区域的零阶矩和一阶矩, 根据式 (8)~式 (9) 利用所得的零阶矩和一阶矩计算 3 个区域质心的坐标, 选择在 X 方向上质心坐标最小的区域, 即为手部区域 (本文研究图像中, 手部区域均在脖子区域的左侧), 保留质心坐标最小的区域, 去除其他区域.

$$m_{00} = \sum_{j=1}^{J}\sum_{i=1}^{I} V(i,j) \tag{5}$$

$$m_{10} = \sum_{j=1}^{J}\sum_{i=1}^{I} i \cdot V(i,j) \tag{6}$$

$$m_{01} = \sum_{j=1}^{J}\sum_{i=1}^{I} j \cdot V(i,j) \tag{7}$$

其中, m_{00}为零阶矩, m_{10}和m_{01}为一阶矩; $V(i,j)$是图像在点(i,j)处的灰度值, I 和 J 分别是图像的宽度和高度.

$$\bar{x} = \frac{m_{10}}{m_{00}} \tag{8}$$

$$\bar{y} = \frac{m_{01}}{m_{00}} \tag{9}$$

手语肤色分割提取的流程图如图 1 所示.

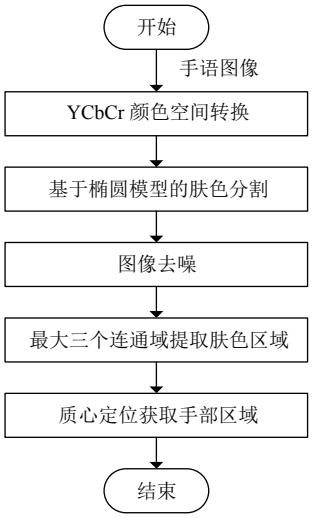

图 1 手语肤色分割提取流程图

手语肤色分割提取的过程结果如图 2 所示.

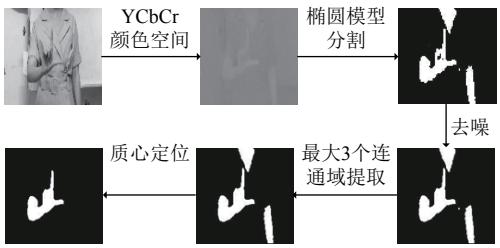

图 2 手语肤色分割提取过程结果

© 中国科学院软件研究所 http://www.c-s-a.org.cn

2 基于改进的 VGG 网络进行手语识别

通过分析 VGG 网络模型的优缺点, 从模型的参数量和计算量方面对 VGG 网络模型进行分析.

2.1 VGG 网络模型介绍

VGG 网络模型在图像特征提取方面具有很明显的优势, 近年来被广泛的用于图像的特征提取[19–21]. 该模型主要是通过增加网络结构的深度来提高网络提取特征的能力, 同时用小的卷积核和小池化核来代替之前的卷积神经网络中的大卷积核和大池化核, 这样既减少了网络结构中的参数量, 又增加了网络中的非线性单元, 提升了神经网络对特征的学习能力. VGG 网络结构如图 3 所示.

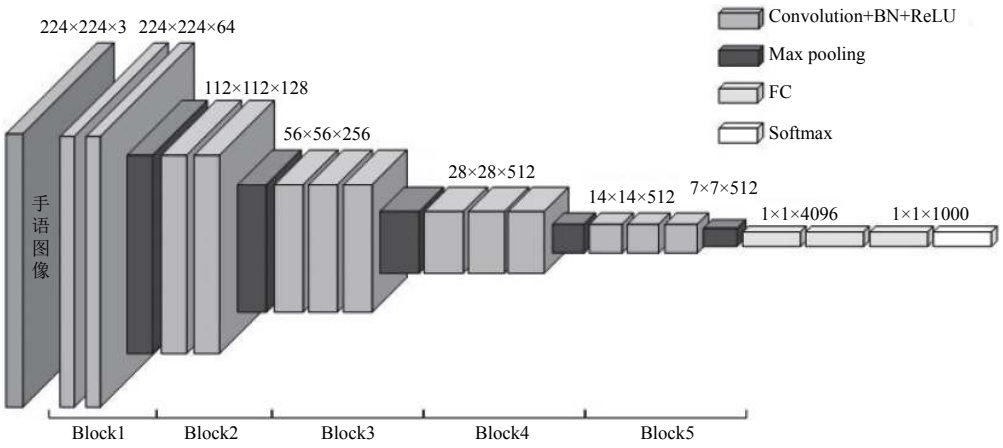

图 3 VGG 网络结构示意图

VGG 网络结构说明如下:

1) 网络的输入是 224×224 的 RGB 图片, 所有图片都经过均值处理.

2) 网络模型中有 5 个 block. 每个 block 内有 2 个或者 3 个卷积层, 共有 13 层卷积; 每个 block 尾部连接一个最大池化层, 用于缩小图片的尺寸, 即压缩输入的特征信息.

3) 网络中总共有 3 个全连接层和一个 Softmax 分类器, 分类器用于对输入的图像进行分类. 在第一个和第二个全连接层后添加了 dropout (随机失活), 这样既可以减少全连接层的计算量, 又避免了网络的过拟合和梯度消散问题.

2.2 改进的 VGG 网络模型

VGG 网络模型在手语识别领域已经取得了不错的成绩[22], 但是 VGG 模型仍存在以下不足之处:

(1) 网络模型的卷积层数太多, 训练模型时计算量大, 损失值的收敛较慢, 且需要大量的数据集;

(2) 通过对 VGG 网络每一层的权重参数量分析可得, VGG 网络训练模型时的参数主要产生于全连接层, 约占整个网络权重参数的 87%, 这就导致了训练网络所需的内存较多.

为了使 VGG 网络模型能够更好地达到手语识别的应用要求, 需要对 VGG 网络结构进行改进, 降低模型所需的存储容量和权重参数量. 对原始的 VGG 网络结构进行如下改进:

(1) 将原来的 13 个卷积层减少到 6 个卷积层, 减少网络对手语图像数据集的需求;

(2) 用两个全连接层代替原来的 3 个全连接层, 并将第一个全连接层的输出节点设为 1024, 第二个全连接层的输出节点设为 26;

(3) 在卷积层和激活函数之间, 我们增加了一个批量归一化 (BN) 层[23], 以提高网络性能和稳定性, 并实现手语图像的准确分类.

BN 是一种有效的逐层归一化的方法, 可以对神经网络中的中间层进行归一化操作, 对于神经网络来说, 令第 l 层的净输入为 $Z^{(l)}$, 经过激活函数后的输出层是 $a^{(l)}$, 如式 (10) 所示.

$$a^{(l)} = f(Z^{(l)}) = f(Wa^{(l-1)} + b) \tag{10}$$

其中, $f(\cdot)$是激活函数, W和 b 是权重和偏置参数.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

为了减少内部协变量偏移问题, 就要使得净输入 $Z^{(l)}$ 的分布一致, 利用数据预处理方法对 $Z^{(l)}$ 进行归一化, 相当于每一层都进行一次数据预处理, 从而加速损失值的收敛速度. 为了提高归一化效率, 一般使用标准归一化, 将净输入 $Z^{(l)}$ 的每一维都归一到标准正态分布, 归一化的公式如式 (11) 所示.

$$\widehat{Z}^{(l)} = \frac{Z^{(l)} - E\left[Z^{(l)}\right]}{\sqrt{var(Z^{(l)}) + \varepsilon}} \tag{11}$$

其中, $E\left[Z^{(l)}\right]$、$var(Z^{(l)})$分别表示当前参数中, $Z^{(l)}$ 的每一维度在整个训练集上的期望、方差, ε为足够小的数.

给定一个包含 K 个样本的小批量样本集合, 第l层神经元的净输入$Z^{(1,l)},\cdots,Z^{(K,l)}$的均值、方差的计算公式分别如式 (12)、式 (13) 所示.

$$\mu_B = \frac{1}{K}\sum_{k=1}^{N} Z^{(k,l)} \tag{12}$$

$$\sigma_B^2 = \frac{1}{K}\sum_{k=1}^{K}(Z^{(k,l)} - \mu_B)\odot(Z^{(k,l)} - \mu_B) \tag{13}$$

为了使归一化操作不对网络的表示能力造成负面影响, 可以通过一个附加的缩放和平移变换改变取值区间, 最后的输出如式 (14) 所示.

$$\widehat{Z}^{(l)} = \frac{Z^{(l)} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}\odot\gamma + \beta \triangleq BN_{\gamma,\beta}(Z^{(l)}) \tag{14}$$

其中, γ、β分别表示缩放和平移的参数向量.

改进的 VGG 网络具体模型结构如图 4 所示, 对比改进前后的网络模型可以看到, 改进后的网络模型卷积层数大大减少, 这就缩短了训练时间. 同时, 改进后的网络中卷积层和池化层依旧是交替出现的, 所以仍保留了图像对缩放、扭曲和位移的不变性和良好鲁棒性的优点.

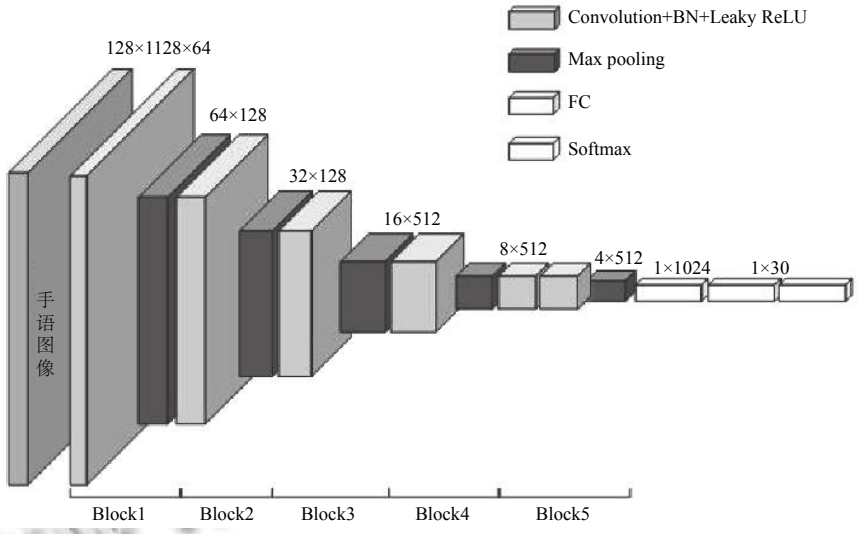

图 4 改进的 VGG 网络结构图

2.3 基于改进的 VGG 的手语识别模型

基于改进的 VGG 网络, 结合手语图像的种类和特点, 构建了识别 26 个英文字母手语的模型, 手语识别流程图如图 5 所示.

(1) 随机从 26 个英文字母手语图像数据集中抽取一定等比例的 26 个英文字母手语图像作为训练样本数据集.

(2) 综合多要素的手语肤色提取分割. 对采集到的手语图像先利用椭圆模型将肤色区域分割出来, 然后再利用最大连通区域和质心定位实现手部区域的分割, 将分割后手语灰度图片的尺寸统一设置为128×128, 并将其作为神经网络的输入.

(3) 模型训练. 利用改进的 VGG 网络提取输入手语图像的特征, 从而构建 26 个英文字母手语图像识别模型.

(4) 模型测试. 手语图像数据集中剩余的手语图像作为测试样本集进行模型测试, 验证模型的准确率.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

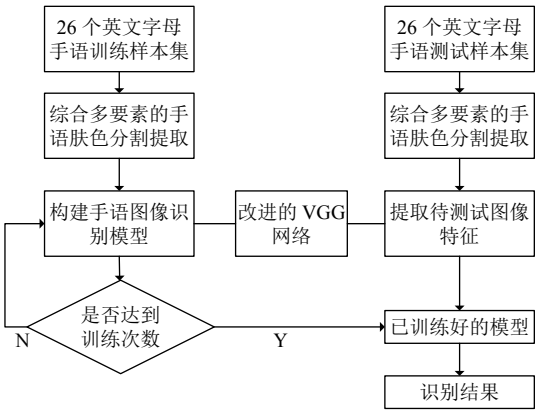

图 5 基于改进的 VGG 的手语识别方法流程图

3 实验与结果

本节主要介绍了手语识别所采用的数据集, 分割算法的有效性验证, 涉及的实验参数设置及网络的对比实验, 实验参数包括批量归一化层 (BN) 的添加、批处理尺寸及学习率的设置.

3.1 数据集介绍

为了验证改进模型的有效性, 本文构建了一个自建手语图像数据集. 自建手势数据集是通过计算机摄像头采集了真人的 26 种不同手语, 共有 10 400 张图像, 手语者穿着类似肤色的衣服, 两侧手臂裸露, 所有手语者均使用右手打手势. 部分手语图像如图 6 所示.

图 6 手语图像数据集

3.2 分割算法有效性验证

为了验证本文提出的手语肤色分割算法的有效性, 我们将本文提出的分割算法和以下 3 种方法进行对比: (1) 基于椭圆模型的肤色分割; (2) 椭圆模型与最大 3 个连通域提取相结合的方法; (3) 椭圆模型与质心定位相结合的方法. 相应的结果图如图 7~图 9 所示, 本文方法的结果图如图 10 所示.

由图 7、图 8 可以看出, 方法 (1) 和方法 (2) 均不能获取单独的手部区域. 由图 9 可以看出, 方法 (3) 只能提取手势图像中最左边一块类肤色区域, 该方法不能实现手部区域的获取. 由图 10 可以看出, 本文的方法对手部区域的获取具有显著效果, 该方法能够从复杂背景中获取单独的手部区域.

图 7 椭圆模型

图 8 椭圆模型与最大 3 个连通域提取相结合的方法

图 9 椭圆模型与质心定位相结合的方法

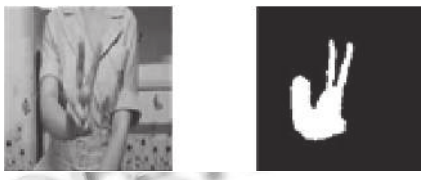

图 10 本文方法

3.3 批量归一化

本实验比较了添加 BN 层和不添加 BN 层的网络训练效果, 对应的损失、准确率随迭代次数的变化如图 11、图 12 所示. 由图 11 可以看出, 添加 BN 层的网络损失值随迭代次数的增加下降较快, 最终趋于稳定; 而未添加 BN 层的网络损失值随迭代次数的增加一直在震荡, 说明添加 BN 层对损失值的下降及稳定具有重要作用. 从图 12 可以看出添加 BN 层的网络准确率明显高于未添加 BN 层的网络准确率高, 说明添加 BN 层有助于获得更高的准确率.

3.4 批处理尺寸及学习率设置

在本实验中, 我们将 batch size 分别设置为 32, 64 和 128, 比较这 3 种条件来选择最适合该模型的 batch size, 不同 batch size 训练的实验结果如图 13、图 14 所

© 中国科学院软件研究所 http://www.c-s-a.org.cn

示. 由图 13 可以看出, 当 batch size = 32 时, 损失值波动幅度远大于其他两种情况, 且梯度下降速率最慢. 当 batch size = 128 时, 损失值波动范围最小. 但是经过一定次数的迭代, batch size 为 64 和 128 的训练情况基本相同. 由图 14 看出, 当 batch size = 32 时, 准确率远大于其他两种情况. 当 batch size = 128 时, 准确率提高较快. 但是经过一定次数的迭代, batch size 为 64 和 128 的训练情况基本相同. 综合考虑, 本实验中选择 64 作为训练的 batch size, 在保证训练速度的同时, 也保证训练模型的泛化能力.

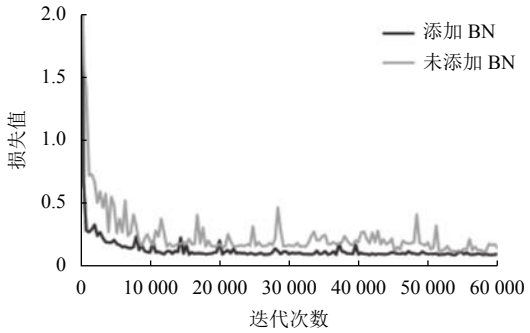

图 11 损失值随迭代次数的变化曲线

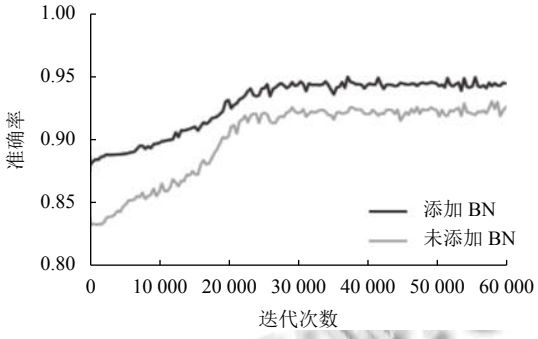

图 12 准确率随迭代次数的变化曲线

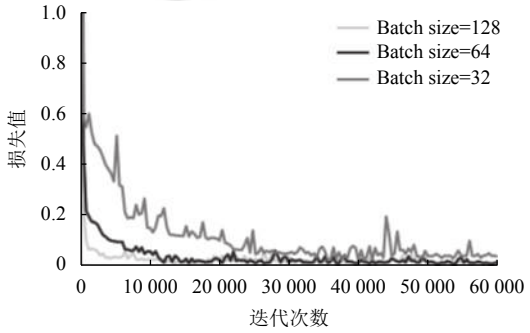

图 13 损失值随迭代次数的变化曲线

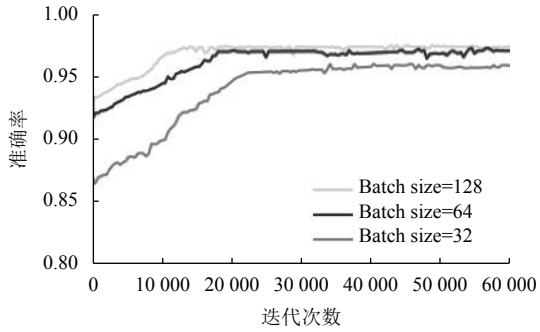

图 14 准确率随迭代次数的变化曲线

本实验将网络的初始学习率 *lr* 设为 0.001, 并且通过指数衰减对学习率进行更新, 衰减系数设为 0.9, 衰减速度设为 1000, 学习率计算公式如式 (15) 所示, 其中 *lr* 为初始学习率, *decay_rate* 为衰减系数, *global_steps* 为当前的迭代次数, *decay_steps* 为衰减速度 (每隔 *decay_steps* 次更新一下学习率).

$$learn\_rate=lr\times decay\_rate^{\frac{global\_steps}{decay\_steps}} \tag{15}$$

3.5 网络的对比试验

通过调整网络中的 block 内的层数来优化网络, 本实验中构建了 4 种网络模型, 如表 1 所示. 由表 1 可以看出, VGG1 网络模型中有 4 个 block, 共有 4 层卷积, 两个全连接层; VGG2 网络模型中有 5 个 block, 共有 5 层卷积, 2 个全连接层; VGG3 网络模型中有 5 个 block, 共有 5 层卷积, 2 个全连接层; VGG4 网络模型 (改进的网络模型) 中有 5 个 block, 共有 6 层卷积, 2 个全连接层. 其中, Conv3 代表卷积层采用3×3的卷积核; Conv3-64 代表该层卷积核的通道数为 64; Max Pooling 代表最大池化层; FC 代表全连接层; FC-1024 代表全连接层的输出节点为 1024.

在实验参数设置相同的基础上, 实验中将讨论 4 种模型训练网络的实验结果.

4 种模型训练网络的实验结果如图 15、图 16 所示. 由图 15 可以看出, 通过比较 VGG1 和 VGG44, 可以发现增加块数来提取更深层次的手语特征, 可以帮助模型较快地实现稳定的收敛. 由 VGG2 和 VGG3 可以发现, 块和卷积层的数量相同时, 增加卷积核的通道数可以提高模型的每个迭代的优化效果最后, 比较 VGG3 和 VGG4 可以发现, 特征深度 (块数) 相同时, 通过增加块内卷积数可以获得更好的特征提取效果. 由图 16 可以看出, VGG4 训练模型的准确率相比其他两种网络模型能够获得较高的识别率, 识别率达到了 97% 以上.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

表 1 卷积网络层配置

层	VGG1	VGG2	VGG3	VGG4
Block1	Conv3-64	Conv3-64	Conv3-64	Conv3-64
	Max Pooling	Max Pooling	Max Pooling	Max Pooling
Block2	Conv3-128	Conv3-128	Conv3-128	Conv3-128
	Max Pooling	Max Pooling	Max Pooling	Max Pooling
Block3	Conv3-256	Conv3-256	Conv3-256	Conv3-256
	Max Pooling	Max Pooling	Max Pooling	Max Pooling
Block4	Conv3-512	Conv3-256	Conv3-512	Conv3-512
	Max Pooling	Max Pooling	Max Pooling	Max Pooling
Block5		Conv3-512	Conv3-512	Conv3-512
				Conv3-512
		Max Pooling	Max Pooling	Max Pooling
FC	FC-1024	FC-1024	FC-1024	FC-1024
FC	FC-26	FC-26	FC-26	FC-26

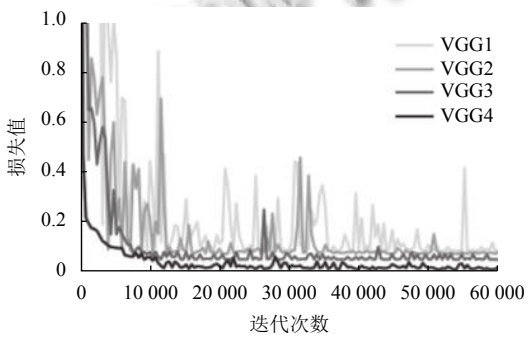

图 15 损失值随迭代次数的变化曲线

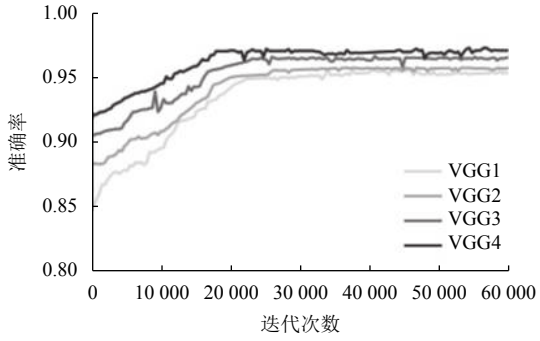

图 16 准确率随迭代次数的变化曲线

4 结论

本文主要研究内容是基于改进的 VGG 网络的手语识别. 在提出实验方案之前, 我们分析了常用的手语特征提取方法的优缺点. 在此基础上, 提出了一种基于综合多要素的手语肤色分割与改进的 VGG 网络结合的手语识别方法. 在该方法中, 根据人体肤色在 YCbCr 空间聚类紧凑的特征构建椭圆模型, 从而对手语图像进行初步分割; 利用中值滤波进行对初步分割后的图形进行平滑处理, 去除肤色区域周围的毛刺或者白点, 然后采用漫水填充算法填充手语区域的空洞, 最后采用基于最大连通域和质心定位的方法手部区域的提取. 本文减少了 VGG 网络模型中的卷积和全连接的层数, 并将批量归一化层添加到网络中. 利用改进后的网络构建识别模型, 识别模型以手部区域的灰度信息为输入, 减少训练网络模型时所需的参数量. 本文提出的方法在保证复杂背景下的手语图像特征提取有效性的同时, 解决了 VGG 网络模型所需数据集大和权重参数量过多等问题, 且保证了手语图像识别的准确性.

参考文献

1 Wu J, Sun L, Jafari R. A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors. IEEE Journal of Biomedical and Health Informatics, 2016, 20(5): 1281–1290. [doi: 10.1109/JBHI.2016.2598302]

2 Kim KW, Lee MS, Soon BR, *et al*. Recognition of sign language with an inertial sensor-based data glove. Technology and Health Care, 2016, 24(Suppl 1): S223–S230.

3 Kumar P, Roy PP, Dogra DP. Independent Bayesian classifier combination based sign language recognition using facial expression. Information Sciences, 2018, 428: 30–48. [doi: 10.1016/j.ins.2017.10.046]

4 Tubaiz N, Shanableh T, Assaleh K. Glove-based continuous Arabic sign language recognition in user-dependent mode. IEEE Transactions on Human-Machine Systems, 2015, 45(4): 526–533. [doi: 10.1109/THMS.2015.2406692]

5 Hrúz M, Trojanová J, Železný M. Local Binary Pattern based features for sign language recognition. Pattern Recognition and Image Analysis, 2012, 22(4): 519–526. [doi: 10.1134/S1054661812040062]

6 Kaur B, Joshi G, Vig R. Indian sign language recognition using Krawtchouk moment-based local features. The Imaging Science Journal, 2017, 65(3): 171–179. [doi: 10.1080/13682199.2017.1311524]

7 Kumar N. Motion trajectory based human face and hands tracking for sign language recognition. Proceedings of the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics. Mathura, India. 2017. 211–216.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

8 Lim KM, Tan AWC, Tan SC. A feature covariance matrix with serial particle filter for isolated sign language recognition. Expert Systems with Applications, 2016, 54: 208–218. [doi: 10.1016/j.eswa.2016.01.047]

9 Pigou L, Dieleman S, Kindermans PJ, *et al*. Sign language recognition using convolutional neural networks. Proceedings of European Conference on Computer Vision. Zurich, Switzerland. 2014. 572–578.

10 Huang J, Zhou WG, Li HQ, *et al*. Sign language recognition using 3D convolutional neural networks. Proceedings of 2015 IEEE International Conference on Multimedia and Expo. Turin, Italy. 2015. 1–6.

11 Hore S, Chatterjee S, Santhi V, *et al*. Indian sign language recognition using optimized neural networks. Proceedings of 2015 International Conference on Information Technology and Intelligent Transportation Systems. Xi'an, China. 2017. 553–563.

12 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1725–1732.

13 Liu X, Chi MM, Zhang YF, *et al*. Classifying high resolution remote sensing images by fine-tuned VGG deep networks. Proceedings of IGARSS 2018–2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia, Spain. 2018. 7137–7140.

14 Gu SS, Ding L. A complex-valued VGG network based deep learing algorithm for image recognition. Proceedings of 2018 Ninth International Conference on Intelligent Control and Information Processing. Wanzhou, China. 2018. 340–343.

15 Ha I, Kim H, Park S, *et al*. Image retrieval using BIM and features from pretrained VGG network for indoor localization. Building and Environment, 2018, 140: 23–31. [doi: 10.1016/j.buildenv.2018.05.026]

16 Amjad A, Griffiths A, Patwary MN. Multiple face detection algorithm using colour skin modelling. IET Image Processing, 2012, 6(8): 1093–1101. [doi: 10.1049/iet-ipr.2012.0167]

17 Rosalina, Yusnita L, Hadisukmana N, *et al*. Implementation of real-time static hand gesture recognition using artificial neural network. Proceedings of the 2017 4th International Conference on Computer Applications and Information Processing Technology. Kuta Bali, Indonesia. 2017. 1–6.

18 Itkarkar RR, Nandi A, Mane B. Contour-based real-time hand gesture recognition for Indian sign language. In: Behera HS, Mohapatra DP, eds. Computational Intelligence in Data Mining. Singapore: Springer, 2017. 683–691.

19 Ye HJ, Han H, Zhu LN, *et al*. Vegetable pest image recognition method based on improved VGG convolution neural network. Journal of Physics: Conference Series, 2019, 1237(3): 032018.

20 Quiroga F, Antonio R, Ronchetti F, *et al*. A study of convolutional architectures for handshape recognition applied to sign language. Proceedings of XXIII Congreso Argentino de Ciencias de la Computación. La Plata, Argentina. 2017. 13–22.

21 Cui RP, Liu H, Zhang CS. Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1610–1618.

22 Wang F, Zhao SS, Zhou XG, *et al*. An recognition–verification mechanism for real-time Chinese sign language recognition based on multi-information fusion. Sensors, 2019, 19(11): 2495. [doi: 10.3390/s19112495]

23 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 448–456.

© 中国科学院软件研究所 http://www.c-s-a.org.cn