

6-20-2023

AR-assisted Sign Language Letter Recognition Method Based on Improved MobileNet Network

Chunhong Liu

School of Computer Science & Technology, Southwest University of Science and Technology, Mianyang 621010, China, 2501649391@qq.com

Song Wang

School of Computer Science & Technology, Southwest University of Science and Technology, Mianyang 621010, China, wangsong@swust.edu.cn

Fupan Wang

School of Computer Science & Technology, Southwest University of Science and Technology, Mianyang 621010, China

Wensheng Tang

School of Computer Science & Technology, Southwest University of Science and Technology, Mianyang 621010, China

See next page for additional authors

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the [Artificial Intelligence and Robotics Commons](#), [Computer Engineering Commons](#), [Numerical Analysis and Scientific Computing Commons](#), [Operations Research](#), [Systems Engineering and Industrial Engineering Commons](#), and the [Systems Science Commons](#)

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

AR-assisted Sign Language Letter Recognition Method Based on Improved MobileNet Network

Abstract

An AR-assisted sign language letter recognition algorithm MS-MobileNet is proposed for the problems of sign language gestures needing to be standardized and low recognition rate. A *multi-scale convolution module is designed to extract the low-level features and enhance the feature extraction ability*. *ELU activation function is used to retain the negative feature information, which combined with a lightweight MobileNet model for the web to improve the recognition accuracy and real-time performance for mobile AR applications*. *Test results show that compared with the original model, the recognition accuracy of MS-MobileNet on the datasets ASL-M, NUS-II and Creative Senz3D is improved by 2.58%, 5.32% and 3.04%, respectively*. *Based on improved network, a WebAR-assisted sign language collaborative interaction system is designed*. After the evaluation test, the average user participation rate reached 8.2 points, and the single recognition time is less than 0.115 s. User's needs for immersive realtime sign language letter interaction is better met.

Keywords

sign language letter recognition; MobileNet; multi-scale convolution; WebAR; collaborative interaction

Authors

Chunhong Liu, Song Wang, Fupan Wang, Wensheng Tang, Yunqiang Pei, Dongsheng Tian, and Yadong Wu

Recommended Citation

Liu Chunhong, Wang Song, Wang Fupan, et al. AR-assisted Sign Language Letter Recognition Method Based on Improved MobileNet Network[J]. Journal of System Simulation, 2023, 35(6): 1308-1321.

基于改进 MobileNet 网络的 AR 辅助手语字母识别方法

刘春宏¹, 王松^{1*}, 王赋攀¹, 唐文生¹, 裴云强¹, 田东生¹, 吴亚东²

(1. 西南科技大学 计算机科学与技术学院, 四川 绵阳 621010; 2. 四川轻化工大学 计算机科学与工程学院, 四川 自贡 643002)

摘要: 针对手语手势姿态待规范、识别率低的问题, 提出一种 AR 辅助手语字母识别算法 MS-MobileNet。设计多尺度卷积模块提取底层特征, 增强网络的特征提取能力; 利用 ELU 激活函数来保留更全面的负值特征信息; 结合适用于 Web 的轻量级 MobileNet 模型, 提高面向移动 AR 应用的手语字母识别准确率和实时性。实验结果表明: MS-MobileNet 在数据集 ASL-M、NUS-II 和 Creative Senz3D 上识别准确率较原模型分别提高了 2.58%、5.32% 和 3.04%。基于 MS-MobileNet 网络设计一套 WebAR 辅助的手语字母协同交互系统, 经评估测试, 用户平均参与度达到 8.2 分, 单次识别耗时低于 0.115 s, 能较好地满足用户沉浸式的实时手语字母交互需求。

关键词: 手语字母识别; MobileNet; 多尺度卷积; WebAR; 协同交互

中图分类号: TP391

文献标志码: A

文章编号: 1004-731X(2023)06-1308-14

DOI: 10.16182/j.issn1004731x.joss.22-0216

引用格式: 刘春宏, 王松, 王赋攀, 等. 基于改进 MobileNet 网络的 AR 辅助手语字母识别方法[J]. 系统仿真学报, 2023, 35(6): 1308-1321.

Reference format: Liu Chunhong, Wang Song, Wang Fupan, et al. AR-assisted Sign Language Letter Recognition Method Based on Improved MobileNet Network[J]. Journal of System Simulation, 2023, 35(6): 1308-1321.

AR-assisted Sign Language Letter Recognition Method Based on Improved MobileNet Network

Liu Chunhong¹, Wang Song^{1*}, Wang Fupan¹, Tang Wensheng¹, Pei Yunqiang¹, Tian Dongsheng¹, Wu Yadong²

(1. School of Computer Science & Technology, Southwest University of Science and Technology, Mianyang 621010, China;

2. School of Computer Science & Engineering, Sichuan University of Science and Engineering, Zigong 643002, China)

Abstract: An AR-assisted sign language letter recognition algorithm MS-MobileNet is proposed for the problems of sign language gestures needing to be standardized and low recognition rate. A multi-scale convolution module is designed to extract the low-level features and enhance the feature extraction ability. ELU activation function is used to retain the negative feature information, which combined with a lightweight MobileNet model for the web to improve the recognition accuracy and real-time performance for mobile AR applications. Test results show that compared with the original model, the recognition accuracy of MS-MobileNet on the datasets ASL-M, NUS-II and Creative Senz3D is improved by 2.58%, 5.32% and 3.04%, respectively. Based on improved network, a WebAR-assisted sign language collaborative interaction system is designed. After the evaluation test, the average user participation rate reached 8.2 points, and the single recognition time is less than 0.115 s. User's needs for immersive real-time sign language letter interaction is better met.

Keywords: sign language letter recognition; MobileNet; multi-scale convolution; WebAR; collaborative interaction

收稿日期: 2022-03-15 修回日期: 2022-05-25

基金项目: 国家自然科学基金面上项目(61872304, 61802320); 四川省自然科学基金(2022NSFSC0961); 西南科技大学博士基金(19zx7144); 西南科技大学素质类教改(青年发展研究)专项(20szjg17)

第一作者: 刘春宏(1995-), 女, 硕士生, 研究方向为移动增强现实、深度学习、人机交互。E-mail: 2501649391@qq.com

通讯作者: 王松(1989-), 男, 讲师, 博士, 研究方向为可视分析与智能交互。E-mail: wangsong@swust.edu.cn

0 引言

手语借助手的姿势、面部表情以及肢体动作搭建了聋哑人与外界沟通的桥梁, 是聋哑人表达思想的主要方式、交流需求的基本保障以及融入社会的重要工具^[1]。其中, 手语字母作为手语的基础组成单元, 因种类的多样性以及普通人未系统地学习手语字母, 容易阻碍聋哑人与普通人之间有效地沟通交流, 因此, 手语字母识别的研究颇为重要^[2]。手语字母识别技术涉及计算机视觉、人机交互等多个交叉领域, 它将人类的意图与智能软硬件系统连接起来。手语字母识别主要分为接触和非接触 2 种方式^[3], 但这些方法存在不同程度的缺陷, 包括投入成本过高、交互不自然、实时性较低等, 进而降低了用户体验。

移动智能设备的普及, 推动了手语字母学习方式的多样化。众多学习方法中, AR^[4-5]是最直观自然的辅助手语字母学习的方式之一。AR 设备^[6]由头戴式和移动增强现实(MAR)两大分支组成, MAR 技术因便于携带且能带给用户较好的 AR 体验, 因此, MAR 能够得到更好的普及^[7]。WebAR 是 MAR 中的一类分支, 它具有可移植、跨平台、轻量等特性。WebAR 环境下的手语字母识别研究, 一方面, 突破了二维平面的限制; 另一方面, 符合用户的自然交互习惯。可通过调整移动端相机镜头的位置为用户提供全方位的观测视角, 有效地解决了普通人学习手语字母过程中存在的手势含义不易理解和手势姿态不标准等问题, 为用户提供更直观的沉浸式手语体验。

1 相关工作

手势识别是 AR、无人驾驶等人机交互领域的主要研究课题。传统的手势识别方法基于可穿戴设备以实现人机交互。Lian 等^[8]提出一种基于肌电传感器的手势识别框架, 用于智能家居手势交互。蔡兴泉等^[9]为提升大场景展览中的便携交互性, 提出一种基于无线数据手套的手势

识别方法。但基于传感器的手势识别因便携性低、体验感欠佳等因素难以推广。随着基于视觉的手势识别方法的广泛应用, Tarvekar 等^[10]提出一种非接触式手势识别应用, 通过 3 种颜色空间检测肤色信息, 肤色阈值分割手势区域, 支持向量机完成手势分类。但基于人工特征提取的方法, 手势识别的准确率和识别速度都有待提升。随着深度学习在图像分类、分割等领域的成功应用, 手势识别的准确性问题逐渐被完善。Barbhuiya 等^[11]利用改进的 AlexNet 网络, 实现静态手语手势识别。Hamid 等^[12]开发了一个基于嵌入式手势识别的应用程序, 将 GoogleNet 和 AlexNet 模型在 Caffe 中进行训练, 用神经运算条测量系统的精度和性能, 使系统满足实时手势识别的需求。Jain 等^[13]提出一种基于 ResNet50 网络的手势识别算法, 其测试准确率较好。伴随着手语手势交流的普及, 在移动或嵌入式设备上设计参数量低、识别速度快的手势识别算法更符合实际的应用需求。Howard 等^[14]提出了一种轻量级网络 MobileNet, 该网络用深度可分离卷积替换标准卷积, 从网络模型的计算复杂度、空间维度和通道数方面进行优化, 保证信息的流畅运行。Liu 等^[15]针对低性能的表面肌电图手势识别, 结合遗传算法优化 MobileNet 模型, 实现低参数高准确率的手势识别应用。He 等^[16]提出一种基于 MobileNetV2 的超高速手势识别算法, 通过网络压缩操作, 提高网络效率。Qian 等^[17]提出基于移动设备的 MobileNetV3 网络分类任务, 相较于其他经典网络, 其准确率更优。Gupta^[18]利用 EfficientNet 等 CNN 模型研发一套手语识别自动化模型, 解决了聋哑人与外界沟通问题。Zhang 等^[19]利用 ShuffleNet 等轻量模型评估毫米波雷达对手势识别的影响, 进而判断其商业价值。Qiang 等^[20]通过减少 SqueezeNet 的结构层数和引入残差结构, 提升了手势检测速度和网络精度。Gu 等^[21]结合分层注意力机制和 MobileNet 网络, 更全面地提取特征信息, 增加

算法的准确性和泛化能力。

基于深度学习的图像处理算法大多用于辅助人机交互研究，而模型质量、硬件性能等因素关系着用户的交互体验。随着深度学习和移动端三维数据的交互需求，面向MAR的手势识别与交互应运而生，虚实融合的设计理念，更直观地展示数据各个维度的信息。Antoshchuk等^[22]将基于视觉的手势识别技术，应用于人机交互界面中，并在AR系统中进行评估。Xiao等^[23]为了解决单模态中用户交互感较低的现状，基于CNN模型的AR多模态交互算法，实现真实手与虚拟模型之间的坐标映射，达到虚实融合的效果。梁欢等^[5]利用传统算法和深度学习方法相结合的方式识别手势，在移动增强现实上用手势交互代替传统的触屏交互。

但上述研究成果很少将基于深度学习的手语字母识别技术与WebAR有效结合，手语字母识别的准确率有待提高等，故本文提出一种面向WebAR环境下基于深度学习的手语字母识别方法。在MobileNet模型的基础上加以改进，包括：

①用多尺度卷积模块提取底层特征；②ELU激活函数代替ReLU激活函数。提出改进后的算法MS-MobileNet，使得在实际应用场景中实现高准确率的手语字母识别目标。以MS-MobileNet算法为核心，设计一套手语字母识别交互系统。系统针对不同需求的人群，设计了3种手语应用场景，用于实现语音、手势和文本信息的自由转换，从而实现手语字母的识别与翻译。经用户评估测试表明，该方案具有低成本、高准确率、性能稳定、协同交互等优势。

2 基于改进的MobileNet手语字母识别算法

手势能辅助听觉和语言能力丧失或受损的患者充分表达自己的想法，因此，手势识别的准确性是人与人之间顺畅沟通交流的关键。基于MS-MobileNet的手语字母识别算法，主要分为3个阶段：手势图像预处理、特征提取和分类预测。图1为基于MS-MobileNet的手语字母识别算法流程图。

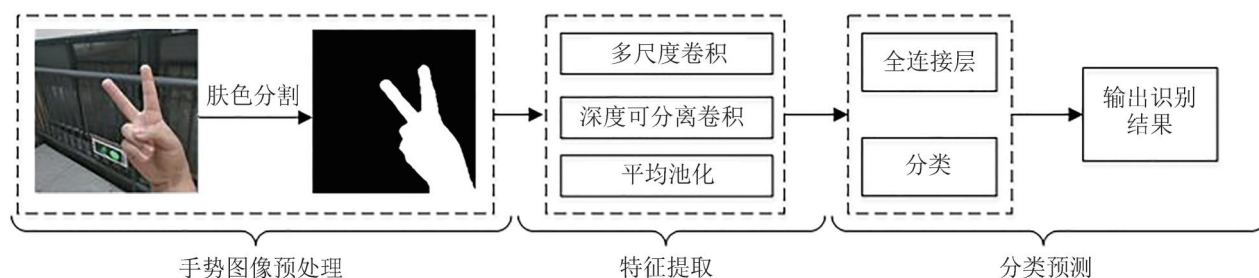


图1 基于MS-MobileNet的手语字母识别算法流程

Fig. 1 Flow chart of sign language letter recognition algorithm based on MS-MobileNet

2.1 图像预处理

通过摄像头采集的RGB图像信息，容易受到光照的干扰，泛化能力弱。相比之下，基于 YC_bC_r 的颜色空间能表现出较好的聚类特性，而且RGB颜色空间转换成 YC_bC_r 颜色空间是呈线性的，计算量较小、分割效果较好。转换公式为^[24]

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

式中： Y 为亮度分量； C_b 和 C_r 分别为蓝色和红色的色度偏移量。

经过颜色空间转换后, 提取 C_r 分量作为手势检测的判断依据, 最后采用 OTSU 算法^[25]做二值化处理, 该方法不易受亮度和对比度的影响, 计算速度快, 还能避免人为地设置阈值, 可提升肤色检测的准确率, 更好地将手势区域和非手势区域分离开, 其流程如图2所示。

2.2 MS-MobileNet 网络的建立

针对复杂的网络模型受在高延迟、高消耗、训练设备受限等因素的限制, 使分类效果有待提升, 以 MobileNet 算法为核心, 分别从卷积方式、激活函数的选择进行改进, 实现响应速度快、准确率高、低参数冗余的手语字母识别算法。改进后的网络模型 MS-MobileNet 的结构图和参数设置分别如图3和表1所示。



图2 手语字母手势肤色分割流程
Fig. 2 Flow chart of sign language letter gesture skin color segmentation

图3 MS-MobileNet 结构图
Fig. 3 MS-MobileNet structure diagram

表1 MS-MobileNet 参数设置
Table 1 MS-MobileNet parameter settings

卷积类型	Filter Shape	Stride	Output Size	卷积类型	Filter Shape	Stride	Output Size
MS-Conv	1×1, 3×3, 5×5	1	224×224×48	Dw Conv 5	3×3×256	1	28×28×256
Dw Conv 1	3×3×48	2	112×112×48	Pw Conv 5	1×1×256×256	1	28×28×256
Pw Conv 1	1×1×48×64	1	112×112×64	Dw Conv 6	3×3×256	2	14×14×256
Dw Conv 2	3×3×64	2	56×56×64	Pw Conv 6	1×1×256×512	1	14×14×512
Pw Conv 2	1×1×64×128	1	56×56×128	Dw Conv 7	3×3×512	2	7×7×512
Dw Conv 3	3×3×128	1	56×56×128	Pw Conv 7	1×1×512×1024	1	7×7×1024
Pw Conv 3	1×1×128×128	1	56×56×128	Dw Conv 8	3×3×1024	1	7×7×1024
Dw Conv 4	3×3×128	2	28×28×128	Pw Conv 8	1×1×1024×1024	1	7×7×1024
Pw Conv 4	1×1×128×256	1	28×28×256				

2.2.1 深度可分离卷积

深度可分离卷积是将标准卷积分解为深度卷积(depth-wise convolution, Dw Conv)和逐点卷积(point-wise convolution, Pw Conv)2部分。深度卷积负责卷积滤波,一个卷积核负责一个输入通道,卷积核的个数等于通道数;逐点卷积负责特征融合,其本质就是普通的标准卷积,但卷积核的大小为 1×1 ^[26]。标准卷积如图4所示,深度可分离卷积过程如图5所示。

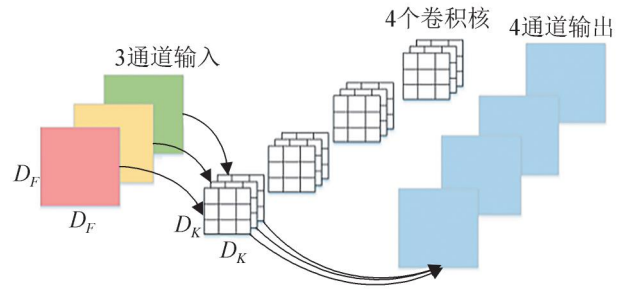


图4 标准卷积

Fig. 4 Standard convolution

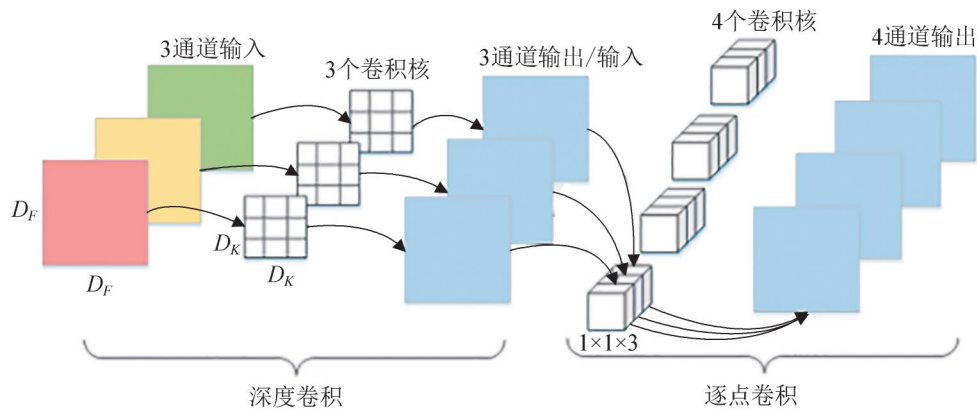


图5 深度可分离卷积

Fig. 5 Depthwise separable convolution

标准卷积过程中输入的特征维度 F 尺寸为 (D_F, D_F, M) ,标准卷积核 K 为 (D_K, D_K, M, N) ,则输出特征维度 G 尺寸为 (D_G, D_G, N) 。其中, $D_F \times D_F \times D_G \times D_G$ 分别表示输入输出特征图各自的宽和高的乘积; $D_K \times D_K$ 表示卷积核的大小; M 表示输入通道数量; N 表示输出通道数量。标准卷积计算公式为^[14]

$$G_{k,l,n} = \sum_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \quad (2)$$

式中: G 为输出特征图; K 为卷积核; F 为输入特征图。则标准卷积的计算量为

$$C_1 = D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \quad (3)$$

深度可分离卷积过程将标准卷积分解为深度卷积和 1×1 逐点卷积2个阶段。深度卷积的计算式为^[14]

$$\hat{G}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \quad (4)$$

式中: $\hat{G}$ 为输出特征图; $\hat{K}$ 为边长为 D_K 的卷积核; F 为输入特征图; k, l 为输出特征图在第 m 通道上的 k, l 坐标; i, j 为第 m 通道的卷积核坐标。

深度卷积核的尺寸为 $(D_K, D_K, 1, M)$;输出特征图的尺寸为 (D_G, D_G, M) ;深度卷积的计算量为

$$C_2 = D_K \cdot D_K \cdot M \cdot D_F \cdot D_F \quad (5)$$

逐点卷积核的尺寸为 $(1, 1, M, N)$;输出特征尺寸为 (D_G, D_G, N) ;逐点卷积的计算量为

$$C_3 = M \cdot N \cdot D_F \cdot D_F \quad (6)$$

因此,深度可分离卷积与标准卷积的计算量的比例为

$$\frac{W_1}{W_2} = \frac{M \cdot D_K^2 \cdot D_F^2 + M \cdot N \cdot D_F^2}{M \cdot N \cdot D_K^2 \cdot D_F^2} = \frac{1}{N} + \frac{1}{D_K^2} \quad (7)$$

由以上公式推导可以看出,在相同纬度特征图的前提下,深度可分离卷积的计算量大幅度减少,则MobileNet网络可以直接部署在移动设备上

进行手语字母图像的识别分类。

2.2.2 基于多尺度卷积的特征提取

手语字母图像存在高低分辨率、复杂背景等特征,用特定且单一的卷积核提取特征信息,限制了网络学习特征的表征范围,使网络学习特征表达能力变弱,扰乱了图像特征的判别度。为了实现低参数量、强鲁棒性的预期模型,大多采用增加网络深度的方式提升识别的准确率和泛化能力。但这种方式对硬件设备要求较高,且伴随着模型的训练时间更长、参数量增加,容易造成梯度爆炸等问题。

针对上述问题,引入多尺度特征提取的思想,在多尺度卷积模块中,卷积核的尺度越小,所表征的目标细节越明显;而尺度越大,网络越深,获得的特征就越抽象。将不同尺度的卷积核级联获取不同尺度的特征图,进而有效地保留了目标的细节信息和抽象信息^[27]。因此,用多尺度卷积(multiscale convolution, MS-Conv)替换MobileNet中标准卷积,构建轻量化手语字母识别模型MS-MobileNet。多尺度卷积模块如图6所示。

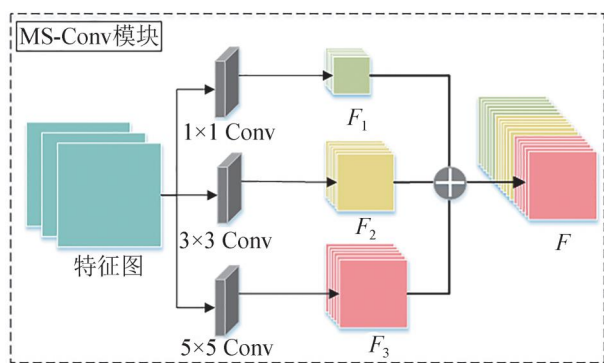


图6 多尺度卷积模块
Fig. 6 MS-Conv module

多尺度卷积模块使用 1×1 , 3×3 和 5×5 尺寸大小的卷积核,每个特征提取通道分别对输入的手语字母手势图像进行卷积运算,多尺度卷积运算表示为

$$\begin{cases} F_i^{1 \times 1} = f\left(\sum_{j=1}^N X_j \cdot K_{ij}^{1 \times 1} + b_i^{1 \times 1}\right) \\ F_i^{3 \times 3} = f\left(\sum_{j=1}^N X_j \cdot K_{ij}^{3 \times 3} + b_i^{3 \times 3}\right) \\ F_i^{5 \times 5} = f\left(\sum_{j=1}^N X_j \cdot K_{ij}^{5 \times 5} + b_i^{5 \times 5}\right) \end{cases} \quad (8)$$

式中: $f(\cdot)$ 为激活函数; N 为输入特征图的子集; X_j 为输入数据层的第 j 个特征图; $K_{ij}^{1 \times 1}$ 为 1×1 滤波器; $b_i^{1 \times 1}$ 为偏置函数; $F_i^{1 \times 1}$ 为多尺度卷积层 1×1 卷积通道的第 i 个输出特征图; $F_i^{3 \times 3}$ 和 $F_i^{5 \times 5}$ 同理可得。

$F_i^{1 \times 1}$ 、 $F_i^{3 \times 3}$ 和 $F_i^{5 \times 5}$ 3个分支的特征图进行通道级联融合运算,得到多尺度特征图,融合后的特征有更多的细节信息,避免了关键特征信息的丢失。特征融合的表达式为

$$F_i = \text{concat}(F_i^{1 \times 1} \oplus F_i^{3 \times 3} \oplus F_i^{5 \times 5}) \quad (9)$$

式中: F_i 为多尺度融合特征图; concat 为连接运算; $\oplus$ 为特征分支融合运算符。

特征映射过程中,多尺度卷积输出通道数依赖于特征提取分支通道的设定。特征融合操作将分支通道进行拼接叠加,提高网络不同尺度的分辨率,增强特征的表达能力。假设3个分支通道分别为 O_α 、 Y_β 和 Z_ω ,特征融合后的通道数为 S_c ,其表达式为

$$S_c = O_\alpha + Y_\beta + Z_\omega \quad (10)$$

在每个卷积层后加入批归一化(BN)层,目的是减少训练时上一层训练参数更新引起的下一层数据分布的改变,BN层可以将数据值调整为标准正态分布,防止梯度消失等,同时加快模型的收敛速度。其表达式为

$$y_i = \gamma \frac{x_i - E}{\sqrt{V + \varepsilon}} + \varphi = \text{BN}_{\gamma, \varphi}(x_i) \quad (11)$$

式中: ε 为分式成立而设定的数值常量,通常是接近0的正数; γ 、 φ 为通过网络自我学习获得2个重构参数; E 为输入 x 的均值; V 为方差。

引入多尺度卷积会增加网络的宽度,但能优化模型的参数量和计算量,提升网络对不同尺度特征的适应性,增加模型的特征感知能力。多尺

度卷积模块的参数量 P 、计算量 Q 的公式为

$$P = \sum_{m=1}^n K_v^2 c_{m-1} c_m \quad (12)$$

$$Q = \sum_{m=1}^n H_m W_m K_v^2 c_{m-1} c_m \quad (13)$$

式中： n 为卷积层； K_v 为滤波器； c_{m-1} 、 c_m 为第 m 个卷积层的输入和输出通道数； H_m 、 W_m 为第 m 个卷积层输入特征的高度及宽度。

2.2.3 激活函数的改进

目前常见的激活函数有 Sigmoid、Tanh、ReLU、PReLU，其中最受欢迎的是 ReLU 激活函数，其图像如图 7 所示。

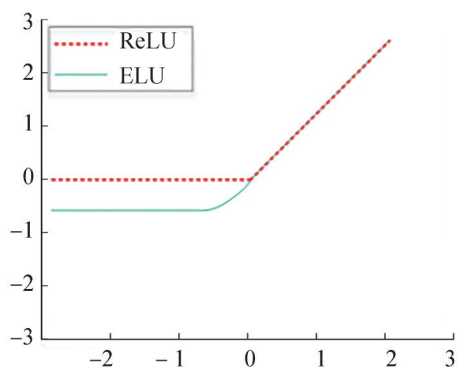


图7 ReLU和ELU激活函数

Fig. 7 ReLU and ELU activation functions

ReLU 数学定义为

$$f(x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (14)$$

当 x 输入为0或正数时，输入等于输出，不存在梯度饱和的情况，能有效地缓解梯度消失的问题，并且是线性计算，速度也有明显的提高。当 x 输入负数时，导数恒为0，不再对数据有激活的作用，也就出现神经元“坏死”的现象，从而导致计算结果不收敛^[28]。

针对 ReLU 激活函数的不足，以及考虑 MS-MobileNet 网络泛化能力，引入了 ReLU 的变体 ELU 激活函数，代替深度可分离卷积中的 ReLU 函数。ELU 图像如图 7 所示，数学定义为 $f(x)$ 和其导数 $f'(x)$ ：

$$f(x) = \begin{cases} x, & x > 0 \\ a(e^x - 1), & x \leq 0 \end{cases} \quad (15)$$

$$f'(x) = \begin{cases} 1, & x > 0 \\ f'(x) + \alpha, & x \leq 0 \end{cases} \quad (16)$$

式中： α 为可调参数，设为0.1~0.3。

由图和公式可知，相较 ReLU 函数而言，当输入 x 为正数时，保持不变；但当 $x < 0$ 时，ELU 函数的导数不为0，保留了负区间的特性，使训练时的梯度保持了较小幅度的更新，避免了神经元死亡，加快了训练的收敛速度。

3 WebAR 辅助的手语字母协同交互系统

为了验证改进算法的合理性和实用性，以 MS-MobileNet 算法为核心，设计了一套聋哑人学习和翻译手语字母的识别交互系统。针对聋哑人对手语日常需求的即时性，以及普通人对手语学习的普适性。将系统搭建于 Web 端，用户只需一台手机或 iPad 等移动设备，即可完成 WebAR 手语字母识别系统的测试和使用。图 8 是 WebAR 辅助的手语字母协同交互系统流程图。

用 3DMax 建模软件绘制出 24 个手语字母的手势骨架模型，Three.js 框架加载模型，进而辅助用户手势姿态的规范化。调用 Web Speech API 接口实现语音合成和识别，实现手势、语音和文本信息的自由转换，有效满足不同用户群体的学习和使用需求。为了提升用户的交互体验，调用移动设备中的陀螺仪装置，通过偏转角速度测量实现不同场景的切换。WebAR 环境下的手语字母识别与协同交互，较好地模拟了聋哑人日常交流的情景，辅助用户学习和理解聋哑手势。

3.1 基于陀螺仪的场景选择

系统基于陀螺仪传感器^[29]通过移动设备倾斜动作角速度的测量，实现主场景之间的切换；通过偏转动作的角速度测量，进入或退出具体主场景。设计场景切换交互功能，目的是贴近用户的使用习惯，丰富系统的交互形式，增加用户操作系统的趣味性。

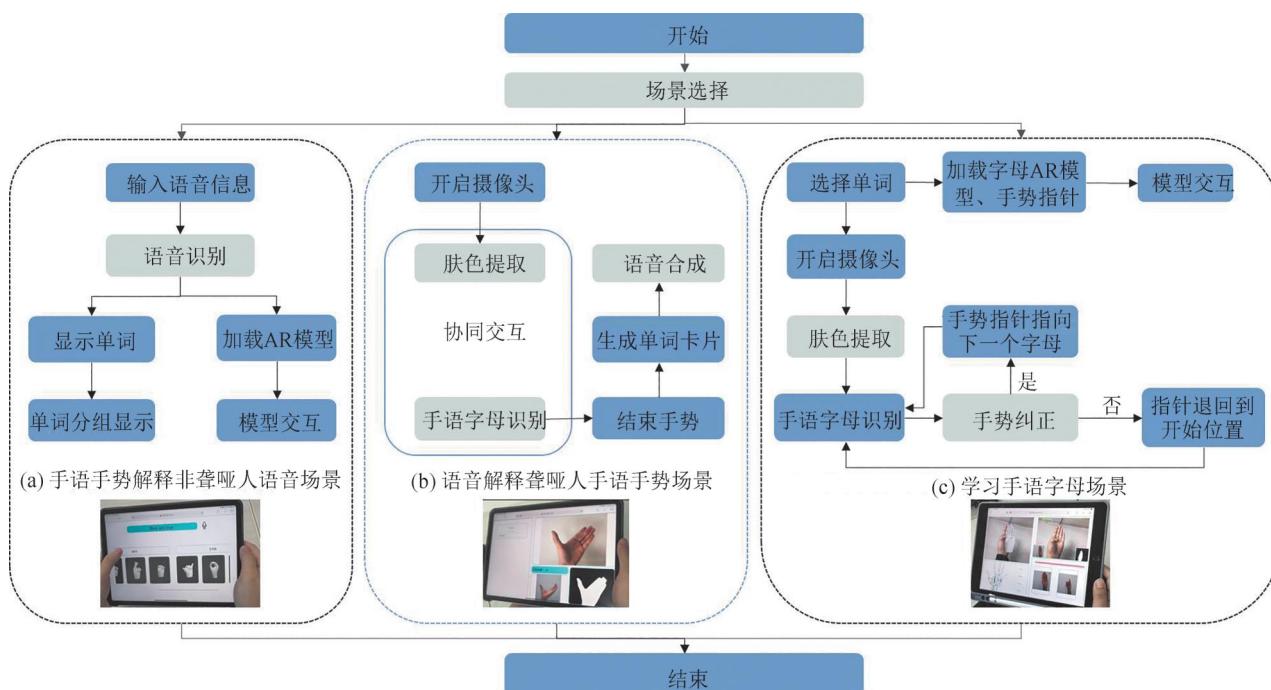


图8 WebAR辅助的手语字母协同交互系统流程

Fig. 8 Flow chart of WebAR-assisted sign language letter collaborative interaction system

3.2 多场景协同交互

(1) 手语手势解释非聋哑人语音场景

该场景主要服务对象为普通人, 其流程如图8(a)所示。通过语音指令加载AR手语字母模型及文本信息, 在AR环境中, 用户与移动设备可实现六自由度下的触控交互, 更直观地学习手语字母。

(2) 语音解释聋哑人手语手势场景

该场景下摄像头捕捉手势图像, 与MS-MobileNet算法训练的手势模型进行预测, 并返回手势识别的结果。当识别到结束手势时, 将其与前结束手势之间的识别结果, 合成单词卡片; 点击卡片播报单词, 实现手势到语音的转换。当健听人与聋哑人之间存在交流阻碍时, 借助该系统复现聋哑人的手势, 理解其所表达的信息, 实现协同交互, 其流程如图8(b)所示。

(3) 学习手语字母场景

该场景由词云可视化、AR手语字母模型、纠正板块组成, 流程图如图8(c)所示。根据维基百科词云可视化统计结果, 用户选择单词后, 触发

相应的AR手语字母模型和纠正功能, 实时地规范手势的姿态; 纠正指针用于判定捕捉当前帧的手势是否正确, 规定时间内, 当前帧的手势预测正确, 纠正指针前进一格, 继续学习后一个手语字母; 反之, 退回到初始指针的位置, 重新学习该单词, 直至学习正确为止。

4 实验分析

4.1 实验环境和评价指标

本实验使用Windows 10操作系统, 显卡为NVIDIA GeForce RTX 2060 SUPER, 8 GB显存, 集成开发环境为PyCharm, 程序编译环境为Python3.8, Tensorflow框架。实验中选择准确率、召回率、精确度、浮点运算次数、参数量等作为实验的评估指标。

网络训练过程中初始学习率设为0.001, 输入图像尺寸均为224×224, batch_size设置为32, 优化器选择Adam, 损失函数使用交叉熵函数, 分类器使用Softmax。

4.2 实验数据集

实际手语字母识别场景中, 识别效果易受外界因素的干扰。为了更好地测试算法的实用性, 本文实验数据采用 Kaggle 的 ASL(american sign language)和部分自建数据集, 统称为 ASL-M 数据集。该数据集的每个数据样本包含不同姿态、位置及光照强度, 共 29 种类别, 其中 26 类分别对应字母 A~Z, 其余 3 类分别为 space、del 和 nothing。选择其中的 24 类字母(除去动态字母 J、Z)和 space 作为本文的数据样本, 共 25 种类别, 每种类别包含 3 200 张图片, 其中 1 200 张为自建数据集, 总计 80 000 张图片, 部分数据图片如图 9 所示。



(a) ASL的部分数据集图



(b) 自建的部分数据集图

图9 ASL-M的部分数据集图
Fig. 9 Partial data set diagram of ASL-M

自建数据集根据 ASL 手语字母样图进行采集, 要求手部距离相机不超过 60 cm, 采集环境包括室内和室外, 灯光和自然光, 目的是贴合实际手势交互的应用场景。为了提高训练模型的鲁棒性, 将实验中的数据集按照 8:2 的比例随机划

分为训练集、测试集, 并使用水平翻转、随机裁剪等方式进行数据扩充操作, 避免发生过拟合的情况。

4.3 性能测试

4.3.1 肤色分割有效性验证

基于 RGB 彩色空间下的图像质量呈现效果受强光和弱光的影响而不同。强光环境下, 设备传感器接收到的光信号较多, 噪声相对较少, 识别率较高; 弱光条件下则相反。基于光照强度、有无肤色分割 2 个条件因素做以下实验, 图 10~11 分别展示了肤色提取过程图和肤色提取性能对比图。图 11 中经肤色提取检测的图像, 一定时间内提取的帧数更多, 判定识别结果相对更稳定。验证了肤色分割算法在本实验中的有效性, 有助于优化网络结构, 增强算法的泛化能力。

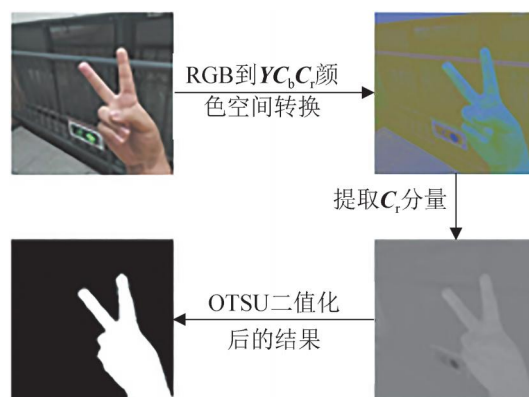


图10 肤色提取过程结果

Fig. 10 Results of skin color extraction process

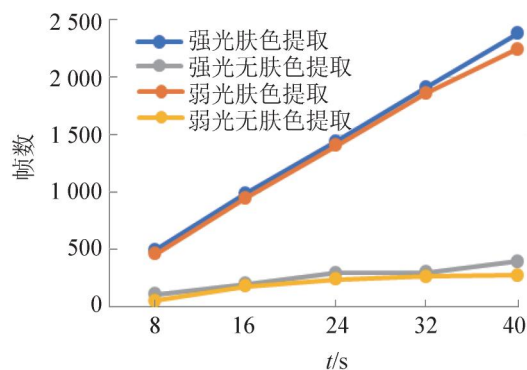


图11 肤色提取性能对比

Fig. 11 Comparison of skin color extraction performance

4.3.2 消融实验

为了验证不同模块对本文算法的影响,在MobileNet算法的基础上进行消融实验分析。从卷积方式、激活函数和二者结合方面改进算法,并设计4组实验进行训练,实验结果如表2所示。

表2 消融实验

Table 2 Ablation experiments		%	
组别	模型	准确率	召回率
1	MobileNet	95.68	94.03
2	MobileNet+ ELU	96.30	96.54
3	MobileNet+MS-Conv	97.13	97.82
4	MobileNet+MS-Conv+ELU	98.26	98.69

从表2可知,第1组实验的准确率和召回率分别为95.68%、94.03%;第2组实验仅将激活函数换成ELU函数,识别效果小幅度增加;第3组实验将标准卷积替换成多尺度卷积,多通道的特征提取方式,提高了手语字母识别准确率;第4组本文的改进算法,准确率提升了2.58%,召回率提升了4.66%。以上实验结果表明,不同分支模块上的改进,准确率都有小幅度的提升,整体改进后的算法手语字母识别效果更好。

4.3.3 对比实验

为了验证本文算法在不同场景任务中的性能效果,在公共数据集NUS-II^[30]和Creative Senz3D^[31-32]进行实验。其中,NUS-II数据集包含10个类别,共2 750个手势;Creative Senz3D数据集只选用静态手势图像,包含11个类别,共1 320个手势,实验结果如表3所示。

表3 在公共数据集上的性能对比

Table 3 Performance comparison on public data sets			%
模型	$A_{\text{NUS-II}}$	$A_{\text{C-S}}$	
本文	92.55	88.41	
MobileNetV1	87.23	85.37	
MobileNetV2	88.46	85.30	
SqueezeNet	88.27	86.03	
EfficientNet	90.63	87.34	
AlexNet	89.31	86.29	

可以看出,本文算法在NUS-II和Creative Senz3D数据集上的准确率分别为92.55%和88.41%,准确率明显高于MobileNetV1;而对比的其余网络整体效果一般,说明数据量偏少的情况下,模型特征感知能力较弱,使手势分类效果有待提高。

手语手势识别任务中,基于构建好的ASL-M数据集进行算法有效性实验,同时设计AlexNet、MobileNetV1、MobileNetV2等9组实验对比,实验结果如表4所示。

表4 不同模型在ASL-M数据集上的性能对比
Table 4 Performance comparison of different models on ASL-M dataset

模型	准确率/%	计算量/百万	参数量/百万
AlexNet	95.21	714	61.10
GoogleNet	90.26	1 504	6.62
ResNet50	94.32	4 111	25.56
MobileNetV1	95.68	575	4.21
MobileNetV2	95.10	314	3.50
MobileNetV3	93.47	312	3.52
ShuffleNet	92.39	148	2.28
SqueezeNet	95.06	823	1.25
EfficientNet	96.37	399	5.30
本文	98.26	580	4.23

由表4可知,本文方法与MobileNetV1相比计算量和参数量偏差不大,但本文方法的识别准确率为98.26%,比MobileNetV1增加了近3%,相较于MobileNetV2、MobileNetV3参数量接近,计算量增加了近270 MB,识别准确率分别增加了3.16%和4.79%,说明多尺度卷积模块,能扩大网络学习特征的范围,提升网络的感知能力,同时结合ELU激活函数,更全面地获取关键信息,使手语字母识别的准确率增加;本文方法和主流的轻量化网络ShuffleNet、SqueezeNet和EfficientNet相比准确率都有相应的增加,权衡参数量、计算量等因素的情况下,本文方法的性能更优;本文方法的参数量和计算量比AlexNet、ResNet50和GoogleNet网络都低,但准确率分别增加了3.05%、3.94%和8%,这是因为本文模型使用深度可分离卷积结构,降低了网络的计算复杂度,

改善了因池化等操作造成的特征信息丢失的情况，提高了手语字母识别的准确性。

根据手语字母应用场景的需要，将本研究算法部署在移动设备上进行测试分析，测试结果如表 5 所示。

表 5 模型的识别效率对比

Table 5 Comparison of model recognition efficiency s		
模型	单次迭代速度	识别时间
MobileNetV1	652	0.132
本文	561	0.115

由表 5 可知，本文算法的收敛速度更快，缩短了模型的训练时长，且识别时间比 MobileNetV1 低 0.017 s，满足移动端实时手语字母的交互需求。

为了进一步分析手语字母识别的泛化能力，对不同类别的手语字母手势进行测试，测试结果如表 6 所示。由表 6 可知，本文算法在测试集上的效果整体较好，但个别手语字母 M 和 N 因相似度过高，使其识别精确度受到了影响。

表 6 不同类别识别性能指标
Table 6 Recognition performance indicators of different categories

字母	精确度	召回率	F1-score	字母	精确度	召回率	F1-score
A	0.99	0.99	1.00	N	0.95	0.97	0.96
B	1.00	1.00	1.00	O	1.00	1.00	0.99
C	1.00	0.99	0.99	P	0.96	0.98	0.98
D	1.00	0.98	0.99	Q	0.98	0.99	0.98
E	0.97	0.97	0.98	R	0.98	1.00	0.99
F	1.00	1.00	1.00	S	1.00	1.00	1.00
G	0.97	0.99	0.99	T	0.98	0.98	0.98
H	0.96	0.95	0.95	U	1.00	0.99	1.00
I	0.97	0.99	0.98	V	1.00	1.00	1.00
K	0.96	0.94	0.96	W	1.00	1.00	1.00
L	1.00	1.00	1.00	X	0.97	0.98	0.98
M	0.94	0.92	0.96	Y	1.00	1.00	1.00

4.4 用户测试

传统视频教学大多采取老师授课、学生听课的教学模式，学生机械化地吸取知识。为了鼓励用户更自主地学习手语字母，在系统中设计了手语字母纠正板块，提高用户的学习动机，系统虚拟仿真界面如图 12 所示。



图 12 系统的虚拟仿真界面
Fig. 12 Virtual simulation interface of system
<http://www.china-simulation.com>

为了验证本文算法的有效性和适用性,邀请65名志愿者参与测试,其中包括研究生26名、本科生39名,且均未经过提前培训。并参考O'Brien等^[33]提出的用户参与度调查,设计一份问卷调查。从集中注意力、实用性、内容吸引力、感知学习性、使用意图和收获程度6个方面进行评估,评估采取10分制的评分模式,评分越高越能反应系统的有效性。要求测试者在使用该系统后完成这份问卷,作为测试该系统的实验依据。图13~14展示在不同教学方式下学习手语字母效果对比图 and 用户测试评估图。

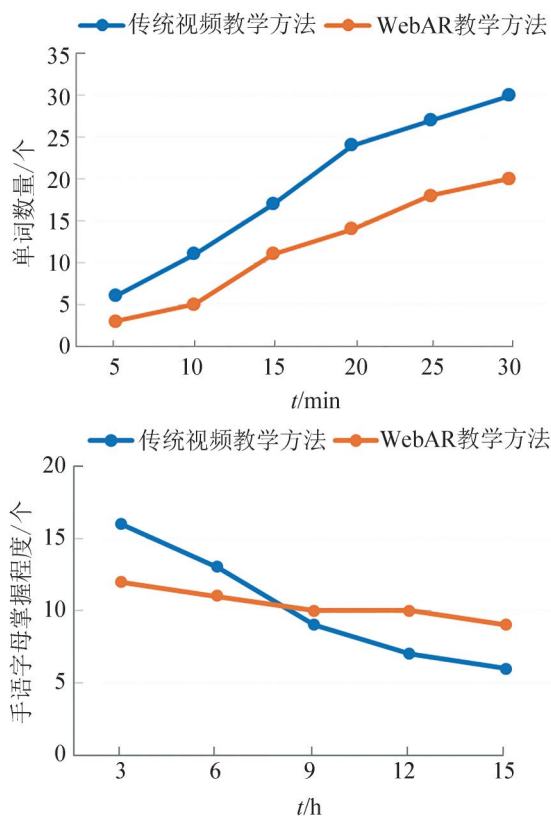


图13 学习单词量和效果对比

Fig. 13 Comparison of number of words and effect of learning

结果表明:相同时间内传统视频教学方法学习的手语字母数量是WebAR教学方法的1.5倍;但随着间隔时长的增加,传统教学方式的遗忘程度越明显,以30 min学习时长为例,其学习效果是WebAR方法的0.7倍。鉴于手势交流是日常生

活中常用的表达形式,熟练掌握每个手语字母是顺畅沟通的基本要素。相比之下,WebAR的教学方式更占优势。由用户评估测试分析可知,用户对该系统在使用意图、收获感受和使用时的专注程度均为9分以上,整体评分达到8分以上。因此,该系统为聋哑手势的初学者,以及需要借助该媒介与聋哑人沟通的用户,提供了一个较好的辅助学习平台,且经用户评估测试验证了系统的实用性、有效性和必要性。

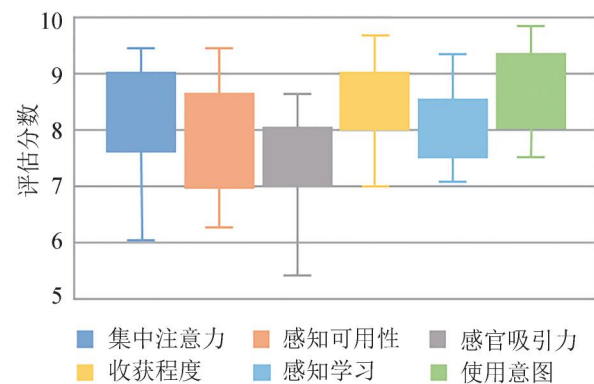


图14 用户参与度评估

Fig. 14 User Engagement Assessment

5 结论

在实际应用场景中,针对手语字母背景复杂化、抗干扰能力弱等问题,可能导致手语字母识别性能不稳定的情况,提出了一种基于MS-MobileNet的手语字母识别算法。将摄像头采集的图像进行预处理,排除多样化背景的干扰,更好地定位手势区域。为了提高手语字母识别在真实场景中的泛化能力,从收敛速度、参数的冗余方面进行优化,并结合多尺度卷积模块,提出改进型网络MS-MobileNet。实验结果可知,本文提出的MS-MobileNet模型比原来的MobileNet模型有更高的准确率。而且轻量化的模型可直接部署在移动设备上,使用更加便捷、泛化能力更好。利用该算法设计的WebAR辅助手语字母识别交互系统,用户体验较好、参与度较高。

未来的工作可以从以下方面展开:①手语字

母识别是孤立词和连续语句识别的基础,但动态手语能较好地传递聋哑人的思想,有助于与外界的沟通;②结合面部表情、肢体运动等非手部信息辅助手语识别,复现实际手语交流场景,推动手语识别智能化;③建立充分详细的手语单词、连续手语数据库,尽早落实手语研究工作服务于实际生活。

参考文献:

- [1] 郝子煜, 阿里甫·库尔班, 李晓红, 等. 基于CapsNet的中国手指语识别[J]. 计算机应用研究, 2019, 36(10): 3157-3159.
Hao Ziyu, Alifu Kuerban, Li Xiaohong, et al. Chinese Finger Language Recognition Using CapsNet[J]. Application Research of Computers, 2019, 36(10): 3157-3159.
- [2] 杨淑莹, 赵敏, 郭杨杨, 等. 基于改进的EfficientDet的手语识别算法[J]. 微电子学与计算机, 2022, 39(2): 84-91.
Yang Shuying, Zhao Min, Guo Yangyang, et al. Sign Language Recognition Algorithm Based on Improved EfficientDet[J]. Microelectronics & Computer, 2022, 39(2): 84-91.
- [3] Köpüklü O, Gunduz A, Kose N, et al. Real-time Hand Gesture Detection and Classification Using Convolutional Neural Networks[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition(FG 2019). Piscataway, NJ, USA: IEEE, 2019: 1-8.
- [4] Amaro N M, Piñero Pérez P Y, Montesinos D F P, et al. Tendencies in Augmented Reality[M]//Piñero Pérez P Y, Bello Pérez R E, Kacprzyk J. Artificial Intelligence in Project Management and Making Decisions. Cham: Springer International Publishing, 2022: 345-362.
- [5] 梁欢, 陈一民, 李德旭, 等. 面向移动增强现实的手势交互方法[J]. 微型电脑应用, 2018, 34(5): 9-13.
Liang Huan, Chen Yimin, Li Dexu, et al. The Method of Hand Gesture Interaction Based on Depth Learning and Hidden Markov Model[J]. Microcomputer Applications, 2018, 34(5): 9-13.
- [6] 裴云强, 吴亚东, 王赋攀, 等. 基于改进L-K光流的WebAR信息可视分析方法[J]. 图学学报, 2020, 41(6): 962-969.
Pei Yunqiang, Wu Yadong, Wang Fupan, et al. IV LKWA: An Information Visual Analysis Tool with Advanced L-K Optical Flow Based WebAR[J]. Journal of Graphics, 2020, 41(6): 962-969.
- [7] Qiao Xiuquan, Ren Pei, Nan Guoshun, et al. Mobile Web Augmented Reality in 5G and Beyond: Challenges, Opportunities, and Future Directions[J]. China Communications, 2019, 16(9): 141-154.
- [8] Lian K Y, Chiu C C, Hong Yongjie, et al. Wearable Armband for Real Time Hand Gesture Recognition[C]//2017 IEEE International Conference on Systems, Man, and Cybernetics(SMC). Piscataway, NJ, USA: IEEE, 2017: 2992-2995.
- [9] 蔡兴泉, 郭天航, 臧坤, 等. 基于无线数据手套的手势识别方法研究[J]. 系统仿真学报, 2014, 26(1): 72-75, 84.
Cai Xingquan, Guo Tianhang, Zang Kun, et al. Gesture Recognition Method Based on Wireless Data Glove[J]. Journal of System Simulation, 2014, 26(1): 72-75, 84.
- [10] Tarvekar M P. Hand Gesture Recognition System for Touch-less Car Interface Using Multiclass Support Vector Machine[C]//2018 Second International Conference on Intelligent Computing and Control Systems(ICICCS). Piscataway, NJ, USA: IEEE, 2018: 1929-1932.
- [11] Barbhuiya A A, Karsh R K, Dutta S. AlexNet-CNN Based Feature Extraction and Classification of Multiclass ASL Hand Gestures[C]//Proceeding of Fifth International Conference on Microelectronics, Computing and Communication Systems. Singapore: Springer Singapore, 2021: 77-89.
- [12] Hamid M F A, Zaman F H K. Hand Gesture Recognition Using Movius Neural Compute Stick[C]//2019 IEEE 9th International Conference on System Engineering and Technology(ICSET). Piscataway, NJ, USA: IEEE, 2019: 510-514.
- [13] Jain A, Sethi A, Vishwakarma D K, et al. Ensembled Neural Network for Static Hand Gesture Recognition[C]//2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). Piscataway, NJ, USA: IEEE, 2021: 1-5.
- [14] Howard A G, Zhu Menglong, Chen Bo, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[EB/OL]. (2017-04-17) [2022-02-26]. <https://doi.org/10.48550/arxiv.1704.04861>.
- [15] Liu Peiyu, Guo Jian, Lu Jiemin, et al. An sEMG Gesture Recognition Algorithm Based on Improved MobileNet [C]//2021 6th International Conference on Computational Intelligence and Applications(ICCIA). Piscataway, NJ, USA: IEEE, 2021: 169-173.
- [16] He Wangpeng, Liao Nannan, Liu Wei, et al. An Ultra-high Speed Gesture Recognition Algorithm Based on MobileNetV2[C]//Neural Computing for Advanced Applications. Singapore: Springer Singapore, 2021:

- 590-602.
- [17] Qian Siying, Ning Chenran, Hu Yuepeng. MobileNetV3 for Image Classification[C]//2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering(ICBAIE). Piscataway, NJ, USA: IEEE, 2021: 490-497.
- [18] Gupta N. Sign Language Recognition Using Diverse Deep Learning Models[C]//International Conference on Artificial Intelligence and Sustainable Engineering. Singapore: Springer Singapore, 2022: 463-475.
- [19] Zhang Kang, Lan Shengchang, Zhang Guiyuan. On the Effect of Training Convolution Neural Network for Millimeter-wave Radar-based Hand Gesture Recognition [J]. Sensors, 2021, 21(1): 259.
- [20] Qiang Baohua, Zhai Yijie, Zhou Mingliang, et al. SqueezeNet and Fusion Network-based Accurate Fast Fully Convolutional Network for Hand Detection and Gesture Recognition[J]. IEEE Access, 2021, 9: 77661-77674.
- [21] Gu Lingyun, Zhang Lin, Wang Zhaokui. Hierarchical Attention-based Astronaut Gesture Recognition: A Dataset and CNN Model[J]. IEEE Access, 2020, 8: 68787-68798.
- [22] Antoshchuk S, Kovalenko M, Sieck J. Gesture Recognition-based Human-computer Interaction Interface for Multimedia Applications[M]//Jat D S, Sieck J, Muyingi H N N, et al. Digitisation of Culture: Namibian and International Perspectives. Singapore: Springer Singapore, 2018: 269-286.
- [23] Xiao Mengting, Feng Zhiquan, Yang Xiaohui, et al. Multimodal Interaction Design and Application in Augmented Reality for Chemical Experiment[J]. Virtual Reality & Intelligent Hardware, 2020, 2(4): 291-304.
- [24] Yusnita L, Hadisukmana N, Wahyu R B, et al. Implementation of Real-time Static Hand Gesture Recognition Using Artificial Neural Network[C]//2017 4th International Conference on Computer Applications and Information Processing Technology(CAIP). Piscataway, NJ, USA: IEEE, 2017: 1-6.
- [25] Otsu N. A Threshold Selection Method from Gray-level Histograms[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.
- [26] 刘文达. 基于单目摄像头的静态手势识别方法研究[D]. 武汉: 湖北工业大学, 2020.
- Liu Wenda. Research of Static Gesture Recognition Methods Based on Monocular Camera[D]. Wuhan: Hubei University of Technology, 2020.
- [27] 景雨, 祁瑞华, 刘建鑫, 等. 基于改进多尺度深度卷积网络的手势识别算法[J]. 计算机科学, 2020, 47(6): 180-183.
- Jing Yu, Qi Ruihua, Liu Jianxin, et al. Gesture Recognition Algorithm Based on Improved Multiscale Deep Convolutional Neural Network[J]. Computer Science, 2020, 47(6): 180-183.
- [28] 蒋昂波, 王维维. ReLU激活函数优化研究[J]. 传感器与微系统, 2018, 37(2): 50-52.
- Jiang Angbo, Wang Weiwei. Research on Optimization of ReLU Activation Function[J]. Transducer and Microsystem Technologies, 2018, 37(2): 50-52.
- [29] 史文策, 许江宁, 林恩凡. 陀螺仪的发展与展望[J]. 导航定位学报, 2021, 9(3): 8-12.
- Shi Wence, Xu Jiangning, Lin Enfan. Development and Prospect of Gyroscope[J]. Journal of Navigation and Positioning, 2021, 9(3): 8-12.
- [30] Pisharady P K, Vadakkepat P, Loh A P. Attention Based Detection and Recognition of Hand Postures Against Complex Backgrounds[J]. International Journal of Computer Vision, 2013, 101(3): 403-419.
- [31] Memo A, Minto L, Zanuttigh P. Exploiting Silhouette Descriptors and Synthetic Data for Hand Gesture Recognition[C]//Smart Tools and Apps for Graphics-Eurographics Italian Chapter Conference. The Eurographics Association, 2015.
- [32] Memo A, Zanuttigh P. Head-mounted Gesture Controlled Interface for Human-computer Interaction[J]. Multimedia Tools and Applications, 2018, 77(1): 27-53.
- [33] O'Brien H L, Cairns P, Hall M. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale(UES) and New UES Short Form[J]. International Journal of Human-Computer Studies, 2018, 112: 28-39.