智能人机交互中第一视角手势表达的一次性学习分类识别

鹿 智 [1] 秦世引 [1, 2] 李连伟 [1] 张鼎豪 [3]

摘 要 在智能人机交互中, 以交互人的视角为第一视角的手势表达发挥着重要作用, 而面向第一视角的手势识别则成为最重要的技术环节. 本文通过深度卷积神经网络的级联组合, 研究复杂应用场景中第一视角下的一次性学习手势识别 (One-shot learning hand gesture recognition, OSLHGR) 算法. 考虑到实际应用的便捷性和适用性, 运用改进的轻量级 SSD (Single shot multibox detector) 目标检测网络实现第一视角下手势目标的快速精确检测; 进而, 以改进的轻量级 U-Net 网络为主要工具进行复杂背景下手势目标的像素级高效精准分割. 在此基础上, 以组合式 3D 深度神经网络为工具, 研究提出了一种第一视角下的一次性学习手势动作识别的网络化算法. 在 Pascal VOC 2012 数据集和 SoftKinetic DS325 采集的手势数据集上进行的一系列实验测试结果表明, 本文所提出的网络化算法在手势目标检测与分割精度、分类识别准确率和实时性等方面都有显著的优势, 可为在复杂应用环境下实现便捷式高性能智能人机交互提供可靠的技术支持.

关键词 智能人机交互, 第一视角, 深度卷积神经网络, 目标检测与分割, 一次性学习手势识别

引用格式 鹿智, 秦世引, 李连伟, 张鼎豪. 智能人机交互中第一视角手势表达的一次性学习分类识别. 自动化学报, 2021, **47**(6): 1284−1301

DOI 10.16383/j.aas.c190754

One-shot Learning Classification and Recognition of Gesture Expression From the Egocentric Viewpoint in Intelligent Human-computer Interaction

LU Zhi[1] QIN Shi-Yin[1, 2] LI Lian-Wei[1] ZHANG Ding-Hao[3]

Abstract In intelligent human-computer interaction (HCI), the expression of gestures with the perspective of the interactive person as the egocentric viewpoint plays an important role, while gesture recognition from the egocentric viewpoint becomes the most important technical link. In this paper, one-shot learning hand gesture recognition (OSLHGR) algorithm under the egocentric viewpoint in complex application scenarios is studied through the cascade combination of deep convolutional neural networks (CNN). Considering the convenience and applicability of practical applications, the improved lightweight SSD (single shot multibox detector) detection network was utilized to achieve rapid and accurate gesture object detection. Furthermore, the improved lightweight U-Net network is used as the main tool to perform pixel-level efficient and accurate segmentation of gesture targets in complex backgrounds. On the basis of U-Net results, a networked algorithm for OSLHGR from the egocentric viewpoint is proposed by using the combined 3D deep neural network. A series of experimental results on the Pascal VOC 2012 dataset and the gesture dataset collected by SoftKinetic DS325 show that the proposed networked algorithm has significant advantages in gesture target detection and segmentation precision, classification accuracy and real-time performance. It can provide reliable technical support for the realization of convenient and high-performance intelligent HCI in complex application environment.

Key words Intelligent human-computer interaction, egocentric viewpoint, deep convolutional neural network, object detection and segmentation, one-shot learning hand gesture recognition (OSLHGR)

Citation Lu Zhi, Qin Shi-Yin, Li Lian-Wei, Zhang Ding-Hao. One-shot learning classification and recognition of gesture expression from the egocentric viewpoint in intelligent human-computer interaction. *Acta Automatica Sinica*, 2021, **47**(6): 1284−1301

收稿日期 2019-10-31 录用日期 2020-01-17
Manuscript received October 31, 2019; accepted January 17, 2020
国家自然科学基金重点项目 (61731001) 资助
Supported by Key Projects of National Natural Science Foundation of China (61731001)
本文责任编委 吴建鑫
Recommended by Associate Editor WU Jian-Xin

1. 北京航空航天大学自动化科学与电气工程学院 北京 100191 2. 东莞理工学院电子工程与智能化学院 东莞 523808 3. 北京航空航天大学电子信息工程学院 北京 100191
1. School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191 2. School of Electrical Engineering and Intelligentization, Dongguan University of Technology, Dongguan 523808 3. School of Electronic Information Engineering, Beihang University, Beijing 100191

面向视觉感知与人机智能交互的工具已逐渐向可穿戴式相机转变, 如 Google Glass、GoPro Hero 和 Narrative Clip 等逐渐成为大众的新宠, 并不断地打入到消费者市场. 这类头戴式相机可用于拍摄运动爱好者的户外探险活动、帮助患有间歇性失忆症的病人记录日常活动、收集用于研究人类行为的

数据和研究以用户为中心的智能人机交互等, 并可以在短时间内记录大量的图像或视频数据. 例如, Narrative Clip 每天能从第一视角拍摄 2 800 多张照片. 因此, 如何帮助人们高效地浏览、搜索和分析第一视角下采集的图像数据, 更好地为人机交互服务变的尤为重要. 随着机器视觉技术的发展, 为解决第一视角下处理图像/视频数据面临的挑战性问题, 包括较差的光照条件和复杂的运动背景等, 提供了新的研究方法.

虽然第一视角下拍摄的视频中包含大量的物体、场景和活动等, 但几乎每帧都包含手这一特定的对象. 这是由于手是我们与物理世界进行交互的主要渠道, 例如, 操作物体、环境感知和人与人之间的肢体交互等. 手总是不断地出现在视野之内, 它的外形和姿势反映出人们正在做什么以及下一步打算做什么. 因此, 手势目标的存在性检测、分割和手势的识别是理解第一视角下人机交互和人人交互的关键性问题. 随着深度学习理论的发展, 大量基于深度 CNN (Convolutional neural network) 的研究工作开始关注于第一视角下手的检测[1]、跟踪[2]、分割[3]和识别[4] 等问题. 然而网络性能在不断提升的同时, 两个挑战性问题阻碍了深度神经网络在便携式移动系统中的应用. 1) 如何在一些特殊的应用领域 (医疗图像、军事卫星图像等) 获取到训练深度神经网络所需的大规模数据集; 2) 计算资源的约束. 通常情况下, 更高的网络性能依赖于大量有标签训练数据对千万级网络参数不断地迭代优化. 而且在便携式移动系统中部署新的网络模型存在许多不可避免的问题, 尤其是在计算资源受限的情况下, 大量的可训练参数、较高的模型计算复杂度和较大的存储空间占用等. 因此, 如何设计轻型高效的网络和如何利用单样本进行高效地分类识别是本文的研究重点.

本文提出了一种基于深度神经网络级联组合解决复杂应用场景中第一视角下的一次性学习手势识别 (One-shot learning hand gesture recognition, OSLHGR) 的算法. 首先, 针对如何快速判定第一视角下手势目标是否出现在相机感受野内的问题, 借助采集的手势目标样本对改进的轻量级 SSD (Single shot multibox detector)[5] 网络进行迁移式强化训练, 一方面可弥补手工制作数据集的不足, 另一方面借助改进 SSD 网络的强实时性的性能优势实现对视频图像序列中手势目标的高效检测. 接着, 在包含手势目标的图像中, 利用改进的 U-Net[6] 模型对复杂背景下的手势目标实施高效精准分割, 以降低无关目标对手势识别结果的影响. 在此基础上, 为实现第一视角下的 OSLHGR, 本文借助于端到端 2D 关系网络 (Relation network, RN)[7], 并将其扩展为处理视频序列输入的 3D 关系网络, 同时采用 3D 残差卷积神经网络 (Residual convolutional neural network) 作为视频数据的特征提取模块. 在对相关类别的大样本数据集进行深度训练的基础上, 使用预训练模型初始化目标网络参数, 提升网络的学习能力, 减少过拟合的风险并加速网络收敛. 在手势目标存在性检测、分割和分类识别的各个阶段, 本文都充分考虑了实际应用对模型高效性与实时性的需求.

本文的主要贡献如下: 1) 在 SSD 目标检测模型的基础上对其进行改进, 以 MobileNetV2[8] 部分网络结构作为 SSD 网络的特征提取模块, 并结合编−解码的思想融合上下文信息, 提出了一种沙漏型的轻量级 SSD 目标检测网络架构. 对比于几种典型的轻量级目标检测模型, 取得了较高的检测性能. 2) 在 U-Net 语义分割模型的基础上, 修改编码网络和解码网络对应层的跳跃连接 (Skip connection), 并使用 1×1 卷积对并置 (Concatenate) 后的特征图进行融合. 改进的 U-Net 相比于轻量级 M2U-Net[9] 在分割精度上有明显的提升, 训练时间略有增加, 但比 U-Net 和 MultiResUNet[10] 模型有明显的速度优势. 3) 创新性地将用于少样本图像识别的 2D 关系网络模型扩展成 3D 关系神经网络并应用到第一视角下的手势识别领域, 通过深度神经网络的自主学习进行特征提取和相似性度量, 降低了网络模型对海量数据的依赖, 首次实现了端到端的 OSLHGR 算法. 4) 为了评估 OSLHGR 算法的分类性能, 使用 SoftKinetic DS325 采集并构建了第一视角下的手势数据集. 该数据集对验证本文提出的第一视角下 OSLHGR 算法的技术路线提供了一个很好的试验基地.

本文的组织结构如下: 第 1 节对智能人机交互与手势表达的优势进行了概述; 第 2 节简要介绍了第一视角手势人机交互的基本环境和约束条件; 第 3 节详细阐述了改进的 SSD 网络结构和基于该模型的手势目标快速检测算法; 第 4 节提出了改进的 U-Net 模型并实现对复杂背景图像中手势目标的高效分割与提取; 第 5 节提出了 3D 关系神经网络并实现了端到端的第一视角下 OSLHGR 算法; 第 6 节对数据集的构建、网络参数的设置和实验结果的性能评价进行了详细介绍; 最后, 对本文的研究工作进行了总结和展望.

1 智能人机交互与手势表达的优势

人机交互 (Human-computer interaction, HCI) 是指人和计算机之间通过某种对话语言, 按照特定的交互方式为完成确定任务而进行的信息交换过程. 在智能化时代, 人与计算机间的交互模式发生

了重大的变化, 不再局限于传统的键盘、鼠标/触控盘和显示屏等交互媒介, 而是逐渐转向集传统交互方式和手势、脑电、眼动和语音等新兴交互方式于一体的多模态交互. 这些交互方式的转变在增加了人机互动的同时, 也获得了更好的享受.

1.1 智能人机交互的应用领域和前沿研究动态

随着计算机、物联网、云计算和人工智能等新兴技术的迅猛发展, 智能人机交互在自动驾驶、医疗、教育、智能机器人、居家和军事等领域有着广泛的应用. 彭玉青等[11] 针对人机交互过程中复杂背景导致手势识别率低、算法鲁棒性差的问题, 提出使用改进的 YOLO[12] 网络完成复杂背景下手势区域的提取并结合 CNN 进行识别. 在医疗领域, Yip 等[13] 提出一种基于眼球追踪眼镜实现手术机械臂的眼动控制界面, 该交互界面允许外科医生通过眼睛观察监视器特定的边缘或角落来控制手术机械臂的运动. 在智能控制机器人方面, Wanluk 等[14] 提出一种专为障碍人群设计的基于眼动跟踪的智能轮椅, 通过对眼球的运动情况进行分析进而控制轮椅的运动. 杨观赐等[15] 提出改进的 YOLO 特征提取算法解决特征提取过程中存在信息丢失的问题, 在隐私情境数据集和服务机器人平台上的实验结果表明了该算法可以较好地识别智能家居环境中涉及隐私的情境. 李昌岭等[16] 提出一种面向未来战场指挥决策的多通道多智能体的人机交互模型, 实现由机器为中心向以人为中心交互的转变, 使得指挥人员和机器间更加自然、无障碍地进行信息交互. 随着技术的不断进步, 未来还会出现更多类型的交互模式, 应用到更多的领域.

1.2 面向人机交互的手势表达的主要方式及人称关系

在人机交互过程中, 手势交互被认为是人与机器间最自然、最便捷的非接触式交互模式. 手势是由人表演的特定姿势或动作来定义, 分为静态和动态手势. 根据相机所处的位置不同, 将基于手势表达的交互方式分为第一视角、第二视角和第三视角下的人机交互[17]. 第一视角下的手势交互由于计算机和表演者的视角是一致的, 计算机看到的也是穿戴者见到的, 可以让计算机更直观地理解操作者的意图. 第二视角下相机是信息接收者, 操作者近距离的面对相机并和计算机进行交互. 对于第三视角下的手势交互, 计算机与操作者的视角不同, 计算机同第三人观察操作者表演手势的视角相同. 操作者可以远离并且背对着相机, 多用于视频监控中. 近年来, 已存在大量的工作对传统视角下的手势识别进行了深入研究. 而随着虚拟现实 (Virtual reality, VR) 和增强现实 (Augmenting reality, AR) 技术的发展, 尤其以 Google Glass 等智能头戴式虚拟现实设备的出现, 第一视角下的手势识别技术也受到了学术界的广泛关注. Hegde 等[18] 为廉价头戴式相机提出了一种可靠且直观的手势交互技术. 在他们的工作中, 首先基于高斯混合模型的手部肤色建模进行前景区域提取, 并利用 Shi-Tomasi 算法计算图像中的特征点, 之后结合 Lukas-Kanade 光流法跟踪前景区域的特征点, 最后对检测到的前景中运动目标进行分类. 随着深度学习理论的发展, 基于深度神经网络的方法也广泛应用于解决第一视角下手势目标的检测、识别等问题. Bambach 等[19] 提出了一种在第一视角下采集的视频中检测和区分不同手势目标的算法, 并在构建的大规模数据集上验证了方法的有效性. Pandey 等[20] 提出使用 MobileNet[21] 作为特征提取的前置网络, 并将 SSD 目标提取网络接在其后, 在移动头戴式显示系统上实现了可靠的手势目标检测和定位.

1.3 第一视角在人机交互中的特点和必要性

随着智能可穿戴设备 (微软 HoloLens、Magic Leap One 等) 的出现并受到越来越多消费者的关注, 第一视角下的人机交互在日常生活中更加普及. 它可以使得人们不会受到任何时间、任何地点和任何环境背景的条件限制, 使用简单定义的手势和头戴式显示系统进行友好交互. 因此, 识别第一视角下的手势动作为我们提供了一种更加自然的与头戴设备中虚拟元素进行交互的模式, 并赋予了人们贴近现实生活的手势导航和控制能力, 建立了与计算机间最直接的交互方式. 在未来智能化的社会中, 第一视角下的手势识别会遍布人们生活的各个角落, 如无人驾驶、智能家居、全息投影、户外运动、机器人控制和体感游戏等. 因此, 第一视角下的手势交互技术需要更多的研究者投入更多的关注, 以解决面临的佩戴者相机抖动、运动模糊、光照变化和背景混杂等问题, 提升人机交互系统在实际应用中的鲁棒性.

1.4 第一视角条件下手势人机交互的优越性

第一视角条件下的手势交互不同于传统视角, 能够感知穿戴者所感知的、看到穿戴者所看到的和理解穿戴者所理解的. 第一视角下的视频是由同一人在连续的时空下录制的, 不需要在环境中放置多个固定的相机, 因此不会受到地理环境、空间和时间的限制, 可以准确记录穿戴者看到的内容, 建立持续、自然的人机交互接口. 此外, 物体和手势是直接呈现在第一视角下的, 不易于被遮挡. 该系统可以识别穿戴者周围的人并了解危险状况, 还可为手

术、运动和娱乐等活动提供帮助. 在自主和可穿戴平台上, 对个人工作空间进行有效地监控也是很多机器人系统的基本要求. 对用户邻近空间内的活动进行可靠、准确和实时的感知也有助于及时做出有意义的决策. 这些都是传统视角条件下的人机交互无法企及的. 因此, 开展第一视角下的手势人机交互具有重要的现实意义.

2　第一视角手势人机交互的基本环境和约束条件

随着智能可穿戴设备逐渐在消费者群体中流行起来, 第一视角下的手势人机交互给人们带来了新的交互方式和交互体验, 摆脱了传统人机交互模式对空间和时间的约束, 拓宽了应用空间.

2.1　面向常规应用的第一视角手势人机交互的基本环境

相比于传统视角下基于手势的智能人机交互, 在第一视角下可以实现全天候的人机交互, 很少会受到时间和空间的制约, 这也促进了第一视角下基于手势人机交互的广泛应用. 如图 1 所示, 展示了在不同光照条件和背景下的第一视角手势人机交互的基本环境. 实际应用中, 用户所处的环境和摄像头固定的位置等因素还是会对第一视角手势人机交互的鲁棒性产生一定程度的干扰. 因此, 如何对人机交互环境中的不利因素进行抑制或消除从而改善智能人机交互系统的整体性能, 是提升良好人机交互体验的关键.

图 1　不同场景下第一视角手势人机交互图示

Fig. 1　HCI demonstration of gestures from the egocentric viewpoint in different scenarios

2.2　实现高性能智能人机交互的第一视角手势表达的约束条件

本文针对第一视角下的 OSLHGR 算法展开研究, 目的是解决复杂背景下依靠单个手势样本的学习实现高性能的智能人机交互. 借助于 SoftKinetic DS325 完成手势数据的采集和测试, 采集示意图如图 2 所示. 深度相机固定在操作者头部正前方的位置, 右半部分由若干线条包围的部分是用于人机交互的区域. 操作者穿戴头部相机的同时, 在规定的区域内执行完预定义的手势动作后, 手离开交互区域并等待下一个动作的执行. 为了对每个动作进行有效地识别, 本文只针对包含单个动作的视频片段进行分类, 并输出相应类别. 计算机再根据输出的类别信息做出相应的响应, 完成一次人机交互过程. 为了使本文所提的算法具有较强的鲁棒性, 采集手势时对表演者手部的配饰品无任何强制要求. 整个手势数据的采集过程是在自然环境中完成的.

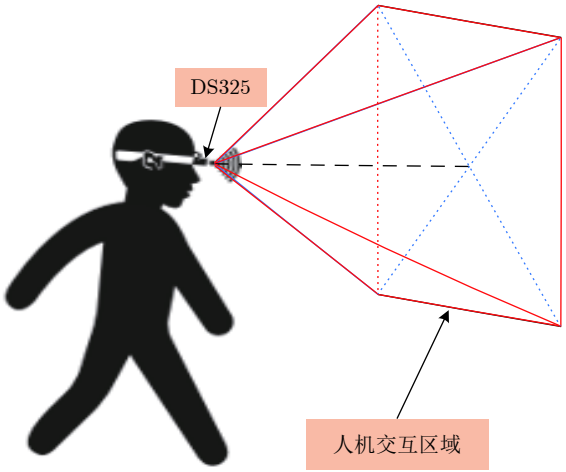


图 2　第一视角下智能人机交互的活动区域图示

Fig. 2　Demonstration of active area of intelligent HCI from the egocentric viewpoint

3　第一视角下的手势目标快速检测

针对头戴式移动设备存在计算能力和存储资源有限的约束问题, 本节在权衡模型精度和计算效率的基础上, 提出一种端到端轻量级目标检测模型, 实现对第一视角下手势目标的快速精准检测. 本节中, 首先对快速检测的要求和工具选择进行分析. 然后对改进的 SSD 网络结构、参数设置和离线监督训练等进行详细地阐述. 最后通过与多种轻量级模型在基准数据集上的检测结果进行对比, 验证了本文提出的检测模型的高效性.

3.1　快速检测的要求与工具选择

第一视角下手势目标的快速精准检测在降低系统响应时间的同时, 还可提升智能人机交互过程中的用户体验. 因此, 本节将针对如何设计高效的目标检测算法, 实现手势目标的快速检测进行研究.

3.1.1　快速检测的性能要求

随着以人为中心的智能交互技术的不断发展, 越来越多的交互设备更加注重用户的体验. 因此, 低延时、高效能的交互系统更受大众青睐. 对于第一视角下基于一次性学习的手势识别算法而言, 实

现较快的手势目标检测速度和较高的召回率有助于提升系统整体的响应时间和分类性能. 随着深度学习理论取得了突破性进展, 基于深度神经网络的目标检测算法在检测性能上要明显优于传统的检测方法[22]. 然而这类算法是通过使用大量训练样本对千万级网络参数不断迭代优化达到较高的检测性能. 因此, 体量大、参数多和复杂性高制约着这些算法在便携式移动系统中的应用. 针对上述问题, 本文通过对 SSD 网络进行轻量化设计, 在实现手势目标快速检测的同时, 大幅降低模型对计算资源的消耗, 实现目标的实时检测.

3.1.2 SSD 网络的检测效能与必要的改进

SSD 是由 Liu 等[5] 提出的一种端到端的目标检测网络模型, 相比于两阶段的目标检测网络 (R-CNN[23], Fast R-CNN[24] 和 Faster R-CNN[25]) 具有明显的速度优势, 而相比于一阶段的 YOLO 网络具有更高的检测精度. 因此, 基于 SSD 在检测速度和精度两方面的性能优势, 本文选取该模型进行轻量化设计. SSD 由两部分组成: 基础网络部分和附加的辅助网络部分. 其中, 基础网络是在 VGG-16[26] 模型的基础上, 用计算量更小的卷积层替换全连接层, 并去除了分类层. 辅助网络是在基础网络部分的基础上新增的 8 个卷积层, 以进一步对基础网络输出的特征图 (Feature map) 进行卷积运算, 并得到多种尺度的特征图. 因此, 可以在多尺度特征图上进行目标类别和位置的预测, 有利于提高目标检测的准确率和增强对低分辨率图像的鲁棒性. 在 SSD 网络中, 输入大小为 300 像素 × 300 像素的图像, 经过一系列的卷积运算, 从基础网络和辅助网络部分选择部分卷积层来实现预测目标边界框的位置和类别. 针对选择的卷积层, 以特征图中每个细胞 (Cell) 单元为中心定义多个包围框 (Default box), 同时用两个卷积层并列的对特征图进行卷积运算, 分别输出预测目标的包围框修正值 (相对于原始包围框的位置偏移量) 和包围框内目标的概率. 基于预测的修正值和原始的包围框, 经过适当变换获得最终的包围框. 训练阶段, 将最终包围框和标注框 (Ground truth) 进行匹配, 计算包括位置误差和置信度误差在内的损失函数, 并使用随机梯度下降算法 (Stochastic gradient descent, SGD) 进行端到端的网络训练. 在预测阶段, 检测模型会生成大量的预测框, 故需使用非极大值抑制 (Non-maximum suppression, NMS) 方法保留具有极大置信度的预测窗口, 即为最终的检测结果.

针对 SSD 网络以 VGG-16 作为基础网络进行特征提取存在着参数多、计算复杂度高和存储消耗大的问题, 改进的 SSD 以轻量级 MobileNetV2 作为基础网络, 并将网络中的标准卷积替换为深度可分离卷积. 此外, 对于 SSD 中不同尺度特征图之间相互独立、低层特征几何细节信息表征能力强而语义信息表征能力弱和高层特征语义表征能力强而几何信息表征能力弱等问题, 本文借鉴文献 [27] 设计出了不对称的沙漏型 SSD 网络结构, 充分融合浅层和深层特征的语义信息, 以此弥补低层次特征语义信息差的问题, 而大多数小目标的检测是依赖于低层次特征图实现的, 因此可提高对小目标的检测和分类精度. 同时将辅助网络中的卷积层替换为 Inception[28] 单元和感受野区块 (Receptive fields block, RFB)[29] 对特征图进行降采样, 增加特征表达能力和鲁棒性. 最后, 受文献 [30] 中采用的基于 SE-Net[31] 注意力机制的启发, 本文将门控 (Gate) 单元加入到网络中的每个预测层, 自适应地选择有用的特征, 进一步增强模型的表达能力. 改进的 SSD 目标检测模型系统架构如图 3 所示. 图中 Depth-wise (DW) 和 Point-wise (PW) 分别表示深度可分离卷积和逐点卷积.

3.1.3 改进 SSD 网络在基准数据集上的性能评价

在第 3.1.2 节的基础上, 按照文献 [5] 中关于目标损失函数的定义, 本文将沿用该损失函数来衡量目标检测的定位损失和目标预测的分类损失, 即

$$L(x,c,l,g)=\frac{1}{N}\left[L_{\text{class}}(x,c)+\alpha L_{\text{loc}}(x,l,g)\right] \tag{1}$$

其中, N 表示和标注框相匹配的默认框的数目. 若 N 等于 0, 表示没有匹配的默认框, 则设置 L 为 0. $L_{\text{class}}(x,c)$ 表示分类损失, 采用交叉熵损失函数, 如式 (2) 所示. $L_{\text{loc}}(x,l,g)$, 如式 (4) 所示. α 表示权值系数, 默认为 1.

$$L_{\text{class}}(x,c)=-\sum_{i\in pos}^{N}x_{ij}^{p}\ln(\hat{c}_i^p)-\sum_{i\in Neg}\ln(\hat{c}_i^0) \tag{2}$$

其中,

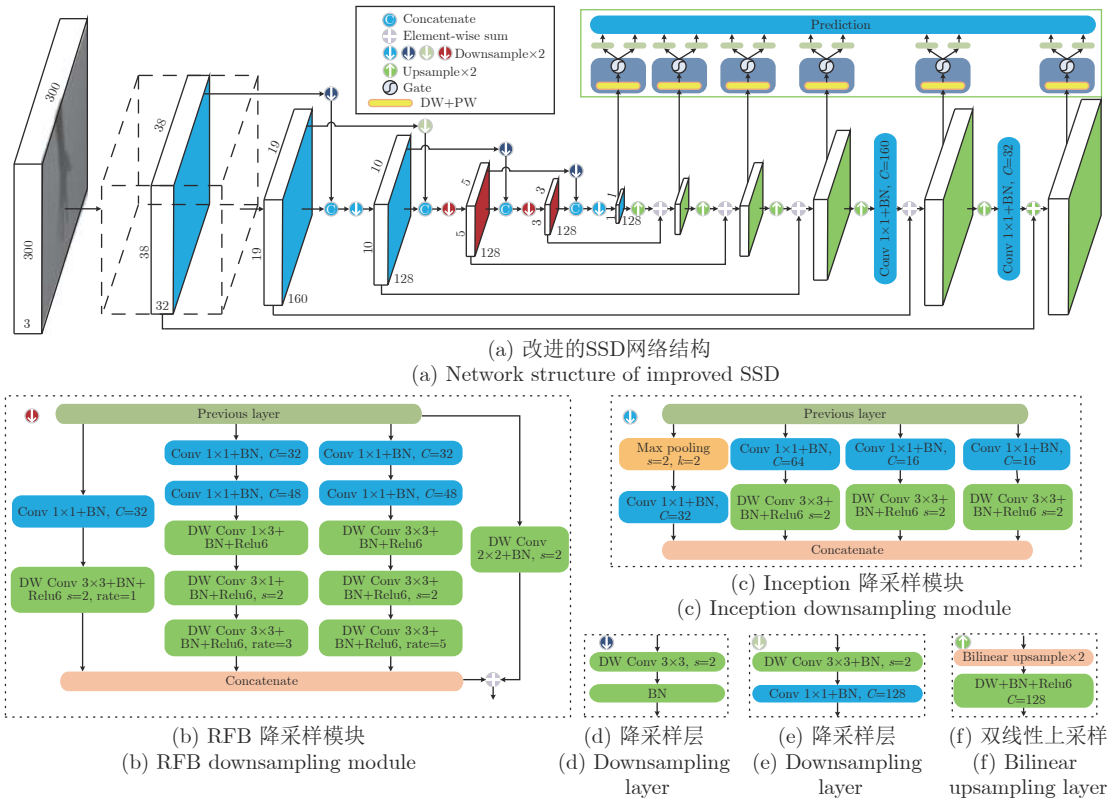
$$\hat{c}_i^p=\frac{\exp(c_i^p)}{\sum\limits_p\exp(c_i^p)} \tag{3}$$

式中, $\hat{c}_i^0$ 表示预测框内没有检测到物体的概率, $\hat{c}_i^p$ 表示第 i 个预测框中的目标是第 p 类的概率.

$$L_{\text{loc}}(x,l,g)=-\sum_{i\in pos}^{N}\sum_{m\in\{cx,cy,w,h\}}x_{ij}^{k}smooth_{L1}(l_i^m-\hat{g}_j^m) \tag{4}$$

其中,

$$smooth_{L1}(x)=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & \text{其他}\end{cases} \tag{5}$$



(a) 改进的SSD网络结构
(a) Network structure of improved SSD

(b) RFB 降采样模块
(b) RFB downsampling module

(c) Inception 降采样模块
(c) Inception downsampling module

(d) 降采样层
(d) Downsampling layer

(e) 降采样层
(e) Downsampling layer

(f) 双线性上采样
(f) Bilinear upsampling layer

图 3　改进的 SSD 目标检测网络架构

Fig. 3　The network architecture of improved SSD object detection

式中, x_{ij}^{k} 取值为 1 或 0, 分别表示第 i 个默认框和第 j 个标注框关于类别 k 是否匹配. l_i^m 表示包围框的预测值, $\hat{g}_j^m$ 则表示标注框的位置参数.

由于改进的 SSD 检测模型是一种新型网络结构, 为防止训练手势目标检测网络时模型过拟合, 通常需使用大规模数据集上的预训练模型初始化目标检测网络的参数, 增强模型的泛化性能. 首先, 在 Pascal VOC 2012 基准数据集上对新型目标检测网络进行充分训练, 并保存检测结果最优的网络模型. 然后, 基于迁移学习[32−33] 的方法使用预训练模型初始化训练手势数据的目标检测网络, 利用 SGD 对损失函数进行优化. 初始学习率设为 10^{-3}, 200 到 250 轮之间, 学习率为 10^{-4}, 250 到 300 轮之间为 10^{-5}, 动量因子为 0.9.

在上述参数设置的基础上, 为了公平地对改进 SSD 模型的效能进行对比分析, 本文以 Pascal VOC 2007 (20 类, 9 963 张图片) 和 VOC 2012 (20 类, 17 125 张图片) 的 trainval 作为训练集, 从头开始训练本文提出的目标检测网络, 并在 test 数据集上进行测试, 实验结果如表 1 所示. 可以看出, 在相似的计算资源约束下, 本文提出的目标检测模型在 VOC 2007 测试数据集上取得了最高的均值平均精度 (Mean average precision, mAP), 达到 73.6%. 尽管相比于原始的 SSD 网络模型, 在检测精度上仍存在差距, 然而改进的 SSD 仅需较少的内存消耗和较低的计算成本. 对比实验结果表明, 改进的 SSD 在计算资源 (模型大小和计算复杂度) 和目标检测精度之间实现了很好的平衡, 更易于满足便携式移动系统的应用需求.

3.2　基于改进 SSD 网络的手势目标快速检测算法

借助上一节中提出的网络结构和在 Pascal VOC 2012 大规模数据集上的离线监督训练. 本文改进的 SSD 网络在目标检测精度和效率上达到了同级别下的较高水平, 基本能够满足对迁移模型的性能需求. 为了充分利用改进 SSD 网络的性能优势, 我们在搭建的实验平台上采集了数百帧第一视角下包含手势目标的图像序列, 并采用 LabelImg 开源标注工具手工制作训练和测试样本集. 在此基础上, 利用迁移学习的策略使用训练集对改进的轻量级 SSD 网络进行微调, 从而实现对第一视角下手势目标的高效和精确检测.

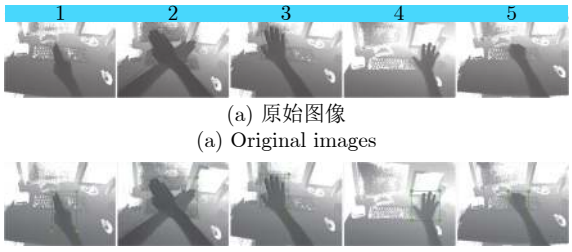
3.2.1　样本的采集与标注

首先, 针对第一视角下手势目标存在性检测的问题, 我们在搭建的数据采集实验平台上采集了 600 帧 (共 10 类手势) 含有手势目标的深度图像,

表 1 轻量级目标检测模型在 VOC 2007 测试集上的检测结果对比 (†表示引用文献 [34] 中的实验结果)
Table 1 Comparison of detection results of lightweight target detection model on VOC 2007 test set († represents the experimental results in [34])

目标检测算法	输入图像大小	训练数据集	测试数据集	mAP (%)	计算复杂度 (M)	参数量 (M)
Tiny YOLO †	416×416	2007 + 2012	2007	57.1	6970	15.12
Tiny SSD[35]	300×300	2007 + 2012	2007	61.3	571	1.13
SqueezeNet-SSD †	300×300	2007 + 2012	2007	64.3	1180	5.50
MobileNet-SSD †	300×300	2007 + 2012	2007	68.0	1140	5.50
Fire SSD[36]	300×300	2007 + 2012	2007	70.5	2670	7.13
Pelee[37]	300×300	2007 + 2012	2007	70.9	1 210	5.98
Tiny DSOD[34]	300×300	2007 + 2012	2007	72.1	1060	0.95
SSD	300×300	2007 + 2012	2007	74.3	34360	34.30
改进的 SSD	300×300	2007 + 2012	2007	73.6	710	1.64

数据采集实验平台见第 6 节. 然后, 使用开源标注工具 LabelImg 对图像中手的位置进行人工标注, 并自动生成对应的 XML 标签数据文件. 标注前后含有手势目标的样本如图 4 所示. 从原始图像中可以看出, 除手外还含有较为复杂的桌面背景, 如显示屏、键盘和鼠标等, 以及出现在第一视角下形状、尺度各异的手势目标都会对手的精确检测带来一定的干扰. 此外, 在标注框内除了手之外, 还有其他对象的干扰. 这说明在检测到手存在的基础上, 需进一步进行精细分割提高手势分类的准确率.



(a) 原始图像
(a) Original images

(b) 手工标注
(b) Manual annotation

图 4 第一视角下手势样本数据的标注结果
Fig. 4 Annotation results of gesture samples from the egocentric viewpoint

3.2.2 基于大样本数据集的强化训练与功能迁移

本节将使用手工标注的 600 幅包含手势目标的图像数据对改进的 SSD 目标检测网络进行深度训练和测试, 其中训练集和测试集按照 5:1 进行随机划分. 由于手工标注的数据集无论是在数据规模和目标类别上都无法和 Pascal VOC 2012 相提并论, 直接用于训练本文提出的目标检测网络模型, 存在过拟合的风险. 鉴于在大规模数据集上提取的浅层视觉特征, 如边缘、纹理、点和线等, 与标注的手势目标数据集之间存在较强的相似性. 因此, 利用第 3.1.3 节在大规模数据集上离线监督训练得到的预训练模型, 并使用迁移学习的策略将预训练模型应用到手势目标检测的任务中, 从而克服手工标注数据的不足, 实现第一视角下手势目标的高效检测. 如图 5 所示, 对比了基于网络模型迁移和 He 等[38] 正态分布两种不同的网络参数初始化策略下, 目标函数随迭代轮次的变化曲线. 从中可以看出, 基于迁移学习的强化训练机制可以使网络的损失函数以更快的速度收敛到较低的值, 实现更高的目标检测和分类预测的性能.

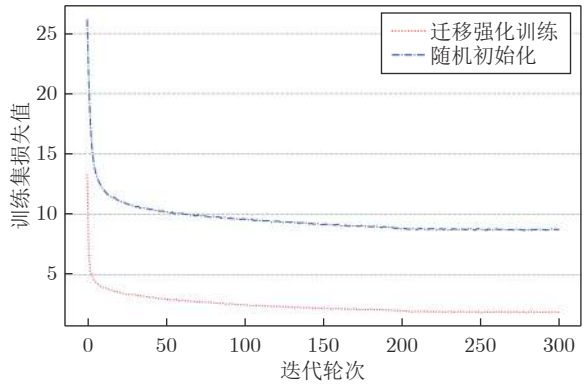
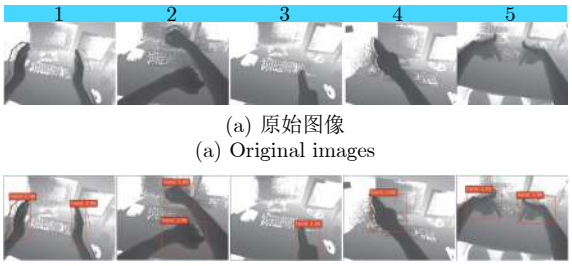


图 5 迁移强化训练和随机初始化两种方式下损失函数变化曲线对比
Fig. 5 Comparison of loss function change curves between transfer reinforcement training and random initialization

3.2.3 第一视角下手势目标的快速检测实验结果和性能评价

在本节中, 首先运用第 3.1.3 节中对改进 SSD 网络进行迁移强化训练而获得的检测模型在 100 帧测试图像上进行手势目标检测. 我们从检测结果中随机选出 5 幅图像, 如图 6 所示. 从中可以看出, 本文改进的轻量级 SSD 对第一视角下采集的包含手的图像, 无论是刚进入到相机感受中尺寸较小的手



(a) 原始图像
(a) Original images

(b) 检测结果
(b) Detection results

图 6　第一视角下改进 SSD 目标检测网络的检测结果

Fig. 6　The detection results of improved SSD target detection network from the egocentric viewpoint

势目标, 还是完全呈现在感受中形状各异的手势目标, 均能精确地进行检测和定位. 这为第一视角下准确高效的判断手在感受野中的存在性提供了重要保障, 也为后续高性能的手势识别奠定了基础.

为了综合衡量本文提出的目标检测算法在采集图像数据集上的检测性能, 我们选择精确率 (Precision) 和召回率 (Recall) 作为目标检测精度的评价指标. 其中, 精确率表示所有检测到的目标中真实手势目标正确检测数所占的比例, 而召回率则表示真实手势目标正确检测数占所有手势目标总数目的比例. 计算表达式分别为

$$P = \frac{T_p}{T_p + F_p} \tag{6}$$

$$R = \frac{T_p}{T_p + F_n} \tag{7}$$

其中, T_p 表示被正确检测为手势目标的帧数, F_p 表示被错误检测为手势目标的帧数, F_n 表示被错误检测为背景的帧数.

将 100 幅测试图像输入训练好的网络模型, 对图像中的手势目标进行检测并记录结果. 当模型输出的预测边界框和测试集中标注的手势目标边界框的交并比 (Intersection over union, IoU) 大于设定阈值时, 检测结果有效. 本文设定阈值为 0.5, 并给出了该阈值下使用预训练模型初始化和随机初始化两种情况下的精确率−召回率变化曲线, 如图 7 所示. 由于只有单类目标, 故 mAP 和 AP 的值相同且均为曲线下方和横纵坐标轴包围区域的面积. 由图中可以看出检测模型在大样本数据集的强化训练下取得了更高的检测性能. 本文在预训练模型初始化网络参数和随机初始化网络参数两种条件下计算 mAP 的值分别为 96.3% 和 94.9%, 这表明改进的 SSD 网络对第一视角下手势目标的检测取得了较高的精度.

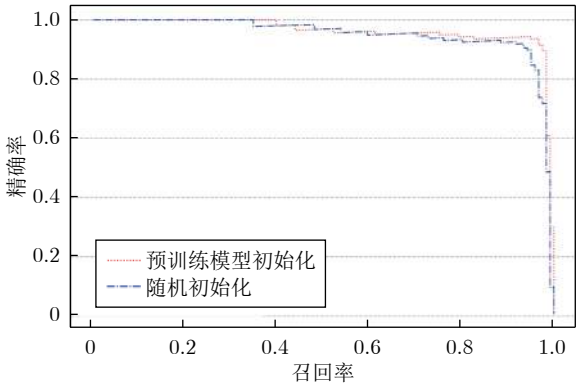


图 7　第一视角下手势目标检测结果的召回率−精确率变化曲线

Fig. 7　Recall and precision curves of gesture target detection results from the egocentric viewpoint

4　基于改进 U-Net 网络的手势目标快速分割与提取

在第 3 节中检测到手势目标存在于相机感受野后, 本节在 U-Net 语义分割模型的基础上提出了一种新的端到端的网络架构, 实现复杂背景下手势目标的高性能分割, 滤除无关目标对手势识别结果的影响. 改进 U-Net 模型结构的设计、参数设置以及深度网络模型的训练在后续小节中分别被详细阐述. 最后对多个语义分割模型在采集图像数据集上的分割结果进行对比, 验证了本文提出的分割模型的高效性.

4.1　改进的轻量级 U-Net 网络模型

随着深度学习理论的发展, 基于深度卷积神经网络的图像分割方法, 如 FCN[39]、U-Net 和 SegNet[40] 等, 相比于传统的分割算法在分割精度上取得了显著地提升. 然而, 这些网络模型普遍存在着参数多、内存消耗大的问题, 无法应用于头戴式移动设备上. 本文在结构简洁、性能更为突出的 U-Net 模型基础上, 设计一种轻量级的全卷积 U 型网络结构用于复杂场景下手势目标的高效分割与提取. 针对 U-Net 存在的问题, 本文提出了三点改进: 1) 将编码端包含大量参数的特征提取网络使用轻量级的 MobileNetV2 替换; 2) 针对编码端和解码端对应层级特征图直接叠加的方式可能存在语义鸿沟的问题, 本文借鉴 MultiResUNet 中使用的 Res path 的思想, 在跳跃连接的支路上通过增加卷积模块来加深低层次卷积层提取深层特征的能力; 3) 在解码端, 对直接叠加的特征图使用 1×1 卷积进行特征融合. 改进后的 U-Net 网络结构如图 8 所示, 其中在编码器部分考虑到模型参数和内存占用等因素选择使用 Mobi-

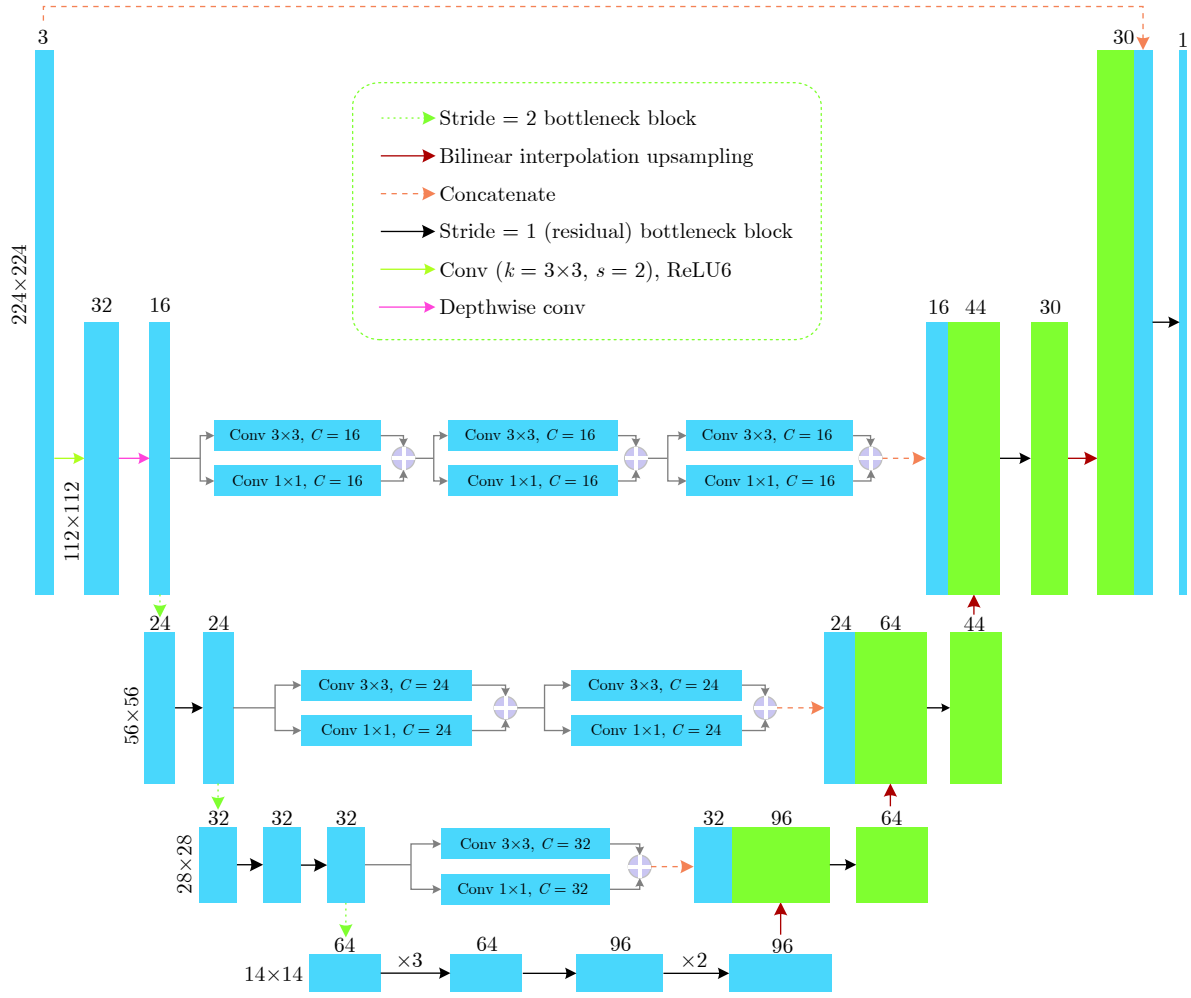


图 8 改进的轻量级 U-Net 网络结构

Fig. 8 Improved lightweight U-Net network architecture

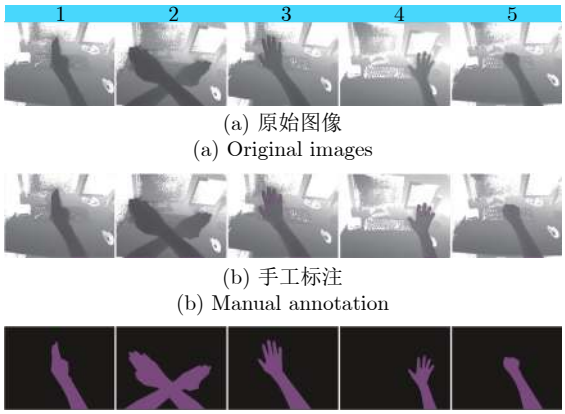
leNetV2 的前 14 层用于特征提取, 同时去除解码器网络中接在双线性插值上采样 (Bilinear interpolation upsampling) 运算后的 2×2 卷积层, 并使用参数量更少的反向残差模块 (Inverted residual block) 将输入通道数减半, 以进一步对网络进行压缩. 对于跳跃连接中使用残差单元的数目是根据编码网络在第一层卷积运算之后进行了三次下采样, 因此在三条跳跃连接支路中从浅层到深层分别增加 3、2 和 1 个残差单元.

改进的轻量级 U-Net 网络模型采用了端到端的对称型网络结构设计, 所有标准卷积都用深度可分离和逐点卷积替代, 极大地降低了网络的参数量和内存消耗. 在采集的图像数据集上对网络模型进行充分训练后, 输入第一视角下采集的原始图像即可快速输出相应大小的分割结果, 因而具备简单、高效的特性. 下一节将对数据的标注、网络模型的深度训练和多种语义分割模型对手势目标的分割结果进行对比分析.

4.2 手势样本数据的标注和网络模型的深度训练

为了对改进的 U-Net 网络模型进行离线监督训练, 我们以第 3 节中使用的 600 幅图像作为网络输入, 并使用 LabelMe 对这些原始图像中的手势目标进行人工标注. 图 9 给出了部分在复杂背景下手势目标的人工标注结果和生成的手势目标区域正样本示例. 图 9(a) 是采集的原始图像, 分别从前五类手势中随机选择的一幅图像. 图 9(b) 是对图像中手势目标人工标注后的结果. 图 9(c) 手轮廓以外的区域表示为背景, 而轮廓以内区域为手势目标的正样本区域. 对改进的 U-Net 网络进行训练之前, 我们将人工标注的 600 幅图像分为两部分: 500 幅图像作为训练集, 100 幅图像用于测试和评估分割模型的性能.

为了对网络参数进行有效地更新和优化, 本文使用二元交叉熵 (Binary cross entropy) 作为损失函数用于度量模型预测输出和期望输出的近似程



(a) 原始图像
(a) Original images

(b) 手工标注
(b) Manual annotation

(c) 手目标区域正样本
(c) Positive sample of hand target area

图 9　第一视角下手势目标轮廓的人工标注结果

Fig. 9　Manual annotation results of gesture target contours from the egocentric viewpoint

度. 同时, 使用自适应矩估计 (Adaptive moment estimation, Adam) 算法对网络参数进行更新, 交叉熵的计算表达式为

$$L = -\sum_{i=1}^{N_n}\left[y^{(i)}\ln(h(x^{(i)})) + (1-y^{(i)})\ln(1-h(x^{(i)}))\right] \tag{8}$$

其中, N_n 表示图像中像素点数, $y^{(i)}$ 表示第 i 个像素的类别, $h(\cdot)$ 使用 Sigmoid 激活函数.

深度神经网络在线训练过程中, 本文对图像进行简单数据增广: 水平翻转、平移变换、旋转变换和缩放变换等, 在不改变训练样本实际数目的同时增加数据的多样性, 使得训练得到的模型泛化性能更好. 为了对不同网络模型的分割结果进行公平地比较, 每个模型都经过 500 轮的迭代训练, 以充分优化网络参数.

4.3　基于改进 U-Net 网络的手势目标快速分割与提取算法

在复杂场景中, 第一视角下采集的手势目标图像包含较多无关的背景干扰, 这对于只通过一次性学习实现高性能手势识别的算法而言会带来较大的挑战. 本文借助于深度学习理论, 利用轻量级 MobileNetV2 网络作为编码端的特征提取模块, 并引入反向残差单元降低卷积层的输入通道数. 此外, 通过在跳跃连接支路上引入不同数目的残差模块, 降低编解码端对应层级特征间的语义鸿沟. 在此基础上, 我们设计出了性能更加优越的目标分割网络模型, 可以实现复杂背景下手势目标的高性能分割. 本文提出的改进 U-Net 网络模型对图像中手势目标分割算法的系统架构如图 10 所示.

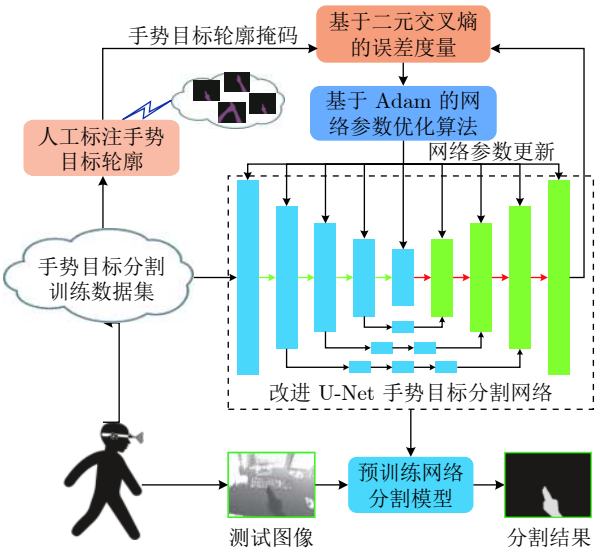


图 10　基于改进 U-Net 的手势目标快速分割和提取算法系统架构

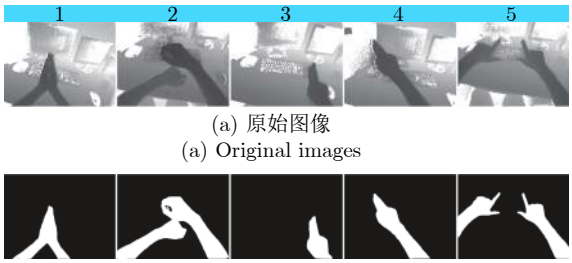
Fig. 10　Architecture of fast segmentation and extraction algorithm of gesture targets based on improved U-Net

在图 10 中, 首先利用建立的数据采集实验平台采集了训练深度神经网络所需的手势样本, 并使用 LabelMe 开源标注工具对采集的包含手势目标图像序列进行人工标注. 将训练数据集和标注后的图像输入网络, 并利用二元交叉熵损失函数计算网络输出结果和人工标注数据间的误差值. 然后使用 Adam 算法对深度网络的参数进行优化, 直至损失函数的值下降到不再变化为止. 在完成对手势目标分割网络模型的训练之后, 实际测试时将获取的手势目标图像输入到训练好的模型, 便可预测输出和输入图像同等大小的手势目标分割结果.

4.4　实验结果与对比分析

为了对测试图像的分割结果有直观的认识, 我们使用第 4.2 节得到的预训练模型对 100 幅测试图像进行预测, 并从分割结果中随机挑选 5 幅图像, 如图 11 所示. 图 11(a) 是原始采集的图像, 图 11(b) 是使用改进的轻量级 U-Net 网络分割的结果. 从中可以看出, 本文提出分割网络模型能够从复杂的图像背景中对手势目标进行有效地分割和提取.

为了综合评估本文提出的网络模型的分割性能, 我们分别对原始的 U-Net 网络、MultiResUNet 网络和轻量级的 M2U-Net 网络在标注的数据集上进行充分训练, 并分别将测试图像输入到训练好的模型中. 并根据式 (9), 计算 100 幅测试图像的平均交并比.



(a) 原始图像
(a) Original images

(b) 分割结果
(b) Segmentation results

图 11 第一视角下改进 U-Net 网络模型的分割结果

Fig. 11 The segmentation results of improved U-Net network model from the egocentric viewpoint

$$IoU = \frac{area(RoI_T \cap RoI_G)}{area(RoI_T \cup RoI_G)} \tag{9}$$

其中, RoI_T 表示不同语义分割模型对原始图像中手势目标的分割结果, RoI_G 表示人工标注的手势目标正样本区域. IoU 的值越大, 说明模型的分割性能越好. 由不同网络模型的分割结果计算出得分如表 2 所示. 同时, 我们还分别给出了各个模型的参数量、计算复杂度和单帧图像的处理时间. 从表 2 中可以看出, 改进的轻量级 U-Net 各项指标均优于原始的 U-Net 网络. 相比于轻量级的 M2U-Net, 本文提出的网络模型以增加较少的计算代价换来模型分割精度的提升. 此外, 对比不同模型处理单幅图像耗费的时间, 可以发现模型的理论计算复杂度和实际的运算时间并不是严格的正相关, 还与网络结构的设计有很大的关系.

表 2 不同网络模型分割结果和模型参数对比

Table 2 Comparison of segmentation results and model parameters of different network models

分割网络模型	输入图像大小	IoU (%)	计算复杂度 (G)	参数量(M)	计算时间 (ms/f)
U-Net	224×224	94.29	41.84	31.03	67.50
MultiResUNet	224×224	94.01	13.23	6.24	119.50
M2U-Net	224×224	93.94	0.38	0.55	36.25
改进的U-Net	224×224	**94.53**	0.52	0.61	53.75

由式 (6) 和式 (7), 我们分别计算了不同网络模型在 100 幅测试图像上手势目标分割结果的召回率和精确率变化曲线, 如图 12 所示. 图 12 中与主对角线交点为 A 的曲线是由本文提出的改进 U-Net 网络模型分割结果计算得到的. 从曲线与坐标轴包围区域的面积和图中标注的 4 个均衡点位置可以看出本文提出的网络结构对第一视角下手势目标的分割结果要明显优于其他几种网络模型.

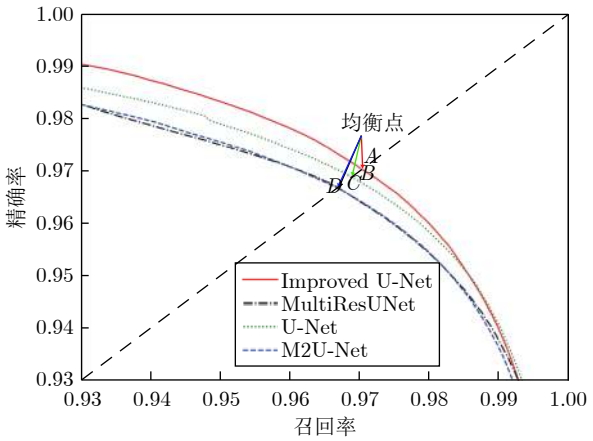


图 12 第一视角下手势目标分割结果的召回率–精确率变化曲线

Fig. 12 Recall and precision curves of gesture target segmentation results from the egocentric viewpoint

为了对比第 3 节与本文提出的轻量级 SSD+U-Net 组合架构和经典的 Mask R-CNN v3[41] 方法在手势目标检测和分割方面的性能, 本节将从定性和定量两个方面阐述所提算法的优越性. 我们使用文献 [41] 中的方法对本文标注的数据进行实验. Mask R-CNN v3 是由 He 等[41] 在 Faster R-CNN[25] 网络模型的基础上增加了目标分割子网络, 在实现有效检测目标的同时输出高质量的目标分割结果. 为了与本文提出的方法进行公平比较, 实验过程中仍使用 500 帧图像进行网络训练和其余 100 帧图像对模型的检测和分割性能进行评价, 测试结果如表 3 所示. 从表 3 中可知, 本文提出的手势目标检测与分割算法相比于经典的 Mask R-CNN v3 方法在保持检测精度无明显损失的情况下取得了较高的分割性能. 此外, 网络参数量大幅度降低也使得模型的检测速度得到了显著提高, 在满足实时检测任务需求的同时也提升了智能人机交互中的用户体验.

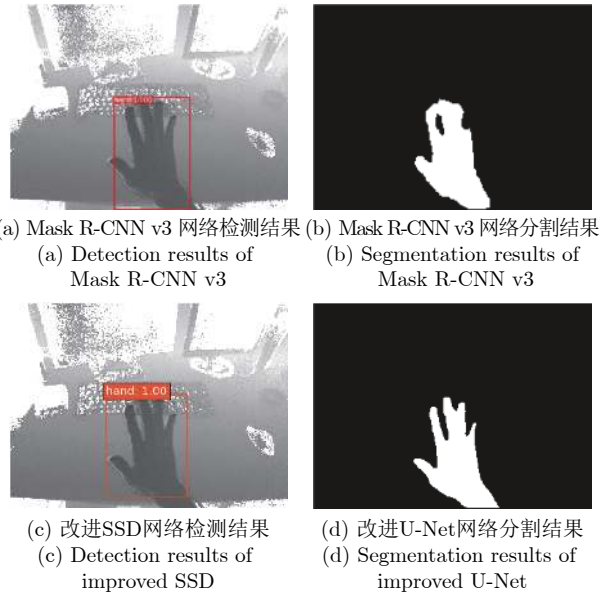
此外, 为了进一步定性地评估本文提出的方法

表 3 本文提出的目标检测和分割方法与 Mask R-CNN v3 的性能对比

Table 3 Performance comparison of the proposed object detection and segmentation method and Mask R-CNN v3

检测与分割算法	输入图像大小	参数量 (M)	mAP (%)	IoU (%)	目标检测时间 (ms/f)
Mask R-CNN v3 (ResNet-101)	300×300	21.20	98.00	83.45	219.73
改进的SSD+U-Net (MobileNetV2)	300×300	2.25	96.30	94.53	31.04

和 Mask R-CNN v3 在手势目标检测与分割结果上的性能, 图 13 中给出了两种方法在本文采集图像数据集上的检测与分割的测试结果. 从图 13 中可以看出, 两种方法均能对手势目标进行精确地检测, 而本文提出的方法在含有手势目标图像上的分割效果要明显优于 Mask R-CNN v3. 因此, 通过对实验结果的定性和定量分析, 可以看出本文提出的轻量级 SSD+U-Net 方法在检测和分割的速度与精度上都能保持在满意的水平.



(a) Mask R-CNN v3 网络检测结果 (b) Mask R-CNN v3 网络分割结果
(a) Detection results of Mask R-CNN v3 (b) Segmentation results of Mask R-CNN v3

(c) 改进SSD网络检测结果 (d) 改进U-Net网络分割结果
(c) Detection results of improved SSD (d) Segmentation results of improved U-Net

图 13　本文提出的 SSD + U-Net 组合方法与 Mask R-CNN v3 检测和分割结果对比

Fig. 13　Comparison of detection and segmentation results between SSD + U-Net and Mask R-CNN v3

5　基于 3D 深度神经网络的一次性学习手势识别

随着手势识别技术的应用在人们日常活动中逐渐增多, 就会不断地出现一些新的赋予不同含义的手势, 这就要求手势识别系统能够快速地对新出现手势进行有效识别. 然而, 在许多实际应用场景中获取大量有标签的训练样本是不切实际的, 这是由于收集或标注数据是非常昂贵和乏味的过程. 本节提出一种新的端到端 3D 关系卷积神经网络用于解决单样本的手势识别问题. 该算法主要是使用 SoftKinetic DS325 采集的第一视角下的深度手势视频数据进行实验.

5.1　3D 深度神经网络结构设计

人类能够通过对单幅样本图像快速学习新类别的原因在于我们大脑中的视觉系统能够非常迅速地提取到图像中物体的显著性特征, 如颜色特征、纹理特征和形状特征等, 再通过比对图像和图像之间的特征差异来实现对目标的识别. 受此启发, Sung 等[7] 提出了一种新颖的关系网络, 通过模拟人类的识别过程来实现对少样本的有效分类. 该网络在训练过程中能够学习一种特征度量方式, 在测试阶段通过计算查询样本和每个新类中单个支撑样本之间的相似度实现对测试图像的分类识别.

在此基础上, 本文将处理图像分类任务的 2D 关系网络修改为解决单样本动态手势分类任务的 3D 关系网络. 本文提出的 3D 关系神经网络系统架构如图 14 所示, 主要包括数据输入单元、特征提取模块、特征相似性度量模块和预测分数输出四个部

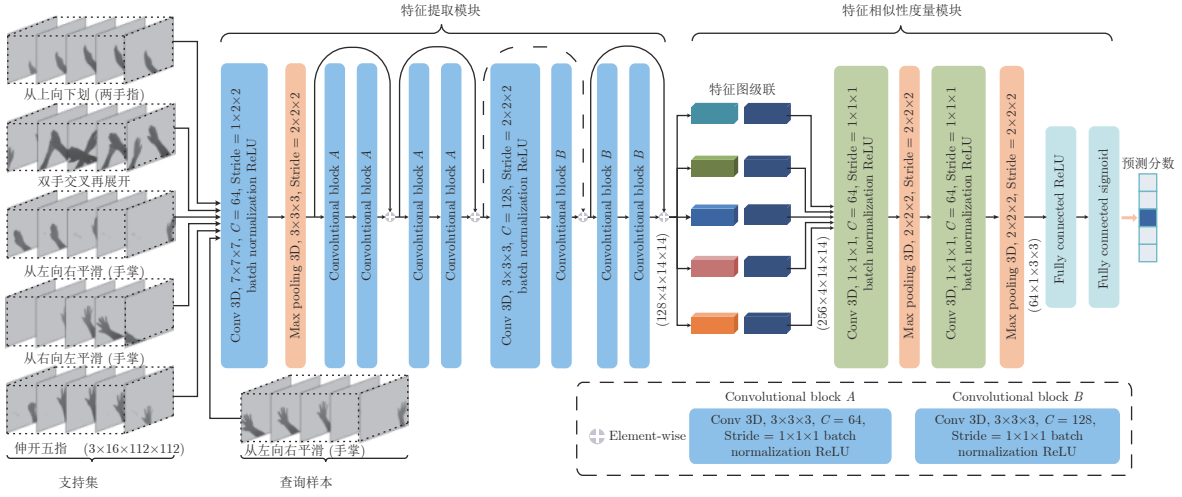


图 14　5-way 1-shot 3D 关系神经网络系统架构

Fig. 14　5-way 1-shot 3D relation neural network system architecture

分. 其中, 输入网络的测试数据集是在第一视角下采集并经过第 3 节和第 4 节处理后的手势视频, 而训练数据是采用文献 [42] 中用于训练网络的 19 类手势数据集, 并确保和测试数据集之间没有相互重叠的类. 关于训练和测试网络模型所用数据划分的细节将在第 5.2 节中详细叙述. 图中特征提取模块使用易于优化和训练的残差网络结构, 本文选择 ResNet9, 并将每一层修改成处理视频序列输入的 3D 卷积运算用于提取时空特征. 特征相似性度量模块是由特征图级联操作和特征度量子网络 (两个 3D 卷积层和全连接层) 组成. 网络的输出是一个值在 [0, 1] 区间内的数, 0 表示查询样本和支撑样本对极不相似, 1 则表示完全相同.

5.2 3D 深度神经网络的训练策略和参数优化

为了便于叙述, 本文首先对用于网络训练和测试的数据进一步细分. 总体上, 我们有三部分数据集: 训练集 (Training set)、支撑集 (Support set) 和测试集 (Testing set). 其中支撑集作为对比学习的样例, 和测试集共享相同的标签. 而训练集的标签则与其他数据集完全不同. 根据测试时的数据结构划分, 本文将具有大量样本的训练集 (Training set) 划分成样本集 (Sample set) 和查询集 (Query set) 两部分来模拟测试时的支撑集和测试集. 对包含 C 个不同的类, 每类有 K 个带标签样本的支撑集, 称为 C-way K-shot (本文只考虑 K =1 的情况) 少样本学习问题. 本文在训练方式上采用和文献 [7] 相同的基于 episode 的策略. 在每次迭代训练网络的过程中, 随机从训练集中选择 C 类且每类包含 K 个带标签的数据样本组成样本集 $S=(x_i,y_i)_{i=1}^{m}$ $(m=C\times K)$, 以及从被选出类别的剩余样本中随机选择一部分样本作为查询集 $Q=(x_j,y_j)_{i=1}^{n}$. 在此基础上对网络反复进行训练, 不断优化模型参数. 此外, 每隔预先设定的迭代次数, 使用支撑集和测试集对当前的网络模型进行测试. 如图 15 所示, 实验中使用的数据集遵循基于 episode 训练方式下的数据划分模式. 图中, 左半部分的元训练集通过多次的 episodes 迭代来模拟一次性学习任务. 在每次迭代过程中, 每类仅含一个正样本 (Positive sample), 用矩形框包围的手势序列表示. 训练阶段, 通过不断地优化网络模型实现对查询样本的最佳分类. 测试阶段, 直接使用优化后的网络模型对测试 episodes 中的查询样本进行预测, 并输出分类结果.

对于单样本学习的手势识别 ($K=1$), 首先将图 14 中的特征提取模块和相似性度量模块分别表示为 f_φ 和 g_ϕ, 并将样本集 S 中的 x_i 和查询集 Q 中的 x_j 输入特征提取网络, 并输出特征图 $f_\varphi(x_i)$ 和 $f_\phi(x_j)$. 然后, 经过特征图级联运算输出特征图 $\mathcal{C}\left[f_\varphi(x_i),f_\varphi(x_j)\right]$, 并输入特征度量模 g_ϕ. 最终经过 Sigmoid 激活函数输出一个值在 [0, 1] 区间内且表示 x_i 和 x_j 相似性程度的关系分数. 因此, 对于 C-way 单样本学习任务而言, 网络输出的关系分数 $s_{i,j}$ 为

$$s_{i,j}=g_\phi\left\{\mathcal{C}\left[f_\varphi(x_i),f_\varphi(x_j)\right]\right\},\ i=1,2,\cdots,C \tag{10}$$

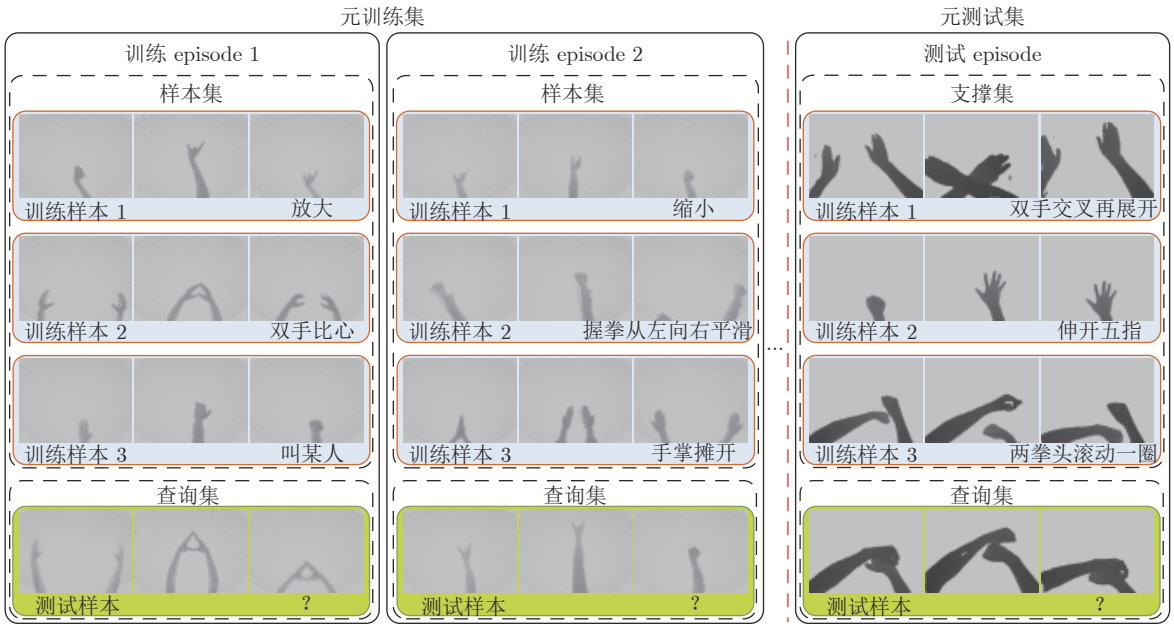


图 15 OSLHGR 任务的数据集划分图示

Fig. 15 Illustration of dataset partitioning for OSLHGR tasks

本文使用均方误差 (Mean square error, MSE) 来衡量预测值和真实值的差异程度, 并使用 Adam 优化器不断地对网络参数进行优化. 其参数优化的决策模型为

$$(\varphi^*, \phi^*) = \arg\min_{\varphi,\phi} \sum_{i=1}^{m}\sum_{j=1}^{n} [s_{i,j}(\varphi,\phi) - I(y_i = y_j)]^2 \tag{11}$$

其中, (φ^*, ϕ^*) 为最优参数集合. $I(\cdot)$ 表示示性函数, 当查询集中样本的标签 y_j 和样本集中样本的预测标签 y_i 相同时, 则 $I(\text{True}) = 1$, 否则为 0.

5.3　一次性学习手势识别算法体系的综合集成与优化

对于第一视角场景下采集的包含手势目标的图像而言, 存在背景复杂、光照变化和头戴式相机抖动等问题. 为了实现高效的 OSLHGR 算法, 首先对手势目标在相机感受野中的存在性进行快速判别, 节省资源的消耗; 其次需对图像中的手势目标进行高效地分割和提取, 从而去除复杂背景对动态手势识别性能的干扰; 最后对分割后的动作序列进行类别判定. 因此, 基于手势目标快速检测、分割和识别的级联组合对第一视角下基于单个手势样本的高效识别是很有必要的.

在 SSD 目标检测模型的基础上进行轻量化设计, 以期在检测精度无明显下降的条件下, 降低模型的计算复杂度并提升目标检测的速度. 此外, 基于迁移学习的方法对改进的 SSD 进行强化训练, 并对第一视角下手势目标进行高效检测. 然后使用改进的 U-Net 模型对图像中的手势目标进行精准分割. 相比于其他图像分割算法, 本文提出的网络模型在分割精度和速度上实现了很好的平衡. 最终将检测和分割处理后的图像序列输入到 3D 关系神经网络, 并输出动态手势的预测结果. 检测、分割和识别相集成的级联组合方案能够满足第一视角下单样本动态手势识别高效性的应用需求, 因此该方案是可行的.

根据第 3 ~ 5 节的研究结果, 本节给出了智能人机交互中第一视角手势表达的一次性学习分类识别算法, 具体见算法 1.

算法 1. 智能人机交互中第一视角手势表达的一次性学习分类识别算法

输入. DS325 采集的深度图像序列 I_{seq}, 支撑集 $S^{'} = \{(x_k, y_k)\}_{k=1}^{C}$.

输出. 动态手势的类别 y.

步骤 1. 依据第 3.2 节提出的训练方法, 获得面向手势目标检测的轻量级 SSD 网络模型 $\mathcal{M}_1$.

步骤 2. For i=1 to length (I_{seq})

步骤 3. 将深度图像序列 I_{seq} 中的每一帧输入到目标检测模型 $\mathcal{M}_1$ 中, 输出图像中存在手势目标的检测结果序列 $I_{\text{det}} = \{I_1, I_2, \cdots, I_n\}$.

步骤 4. End For

步骤 5. 依据第 4.2 节使用的深度神经网络训练方法, 获得面向手势目标分割的轻量级 U-Net 网络模型 $\mathcal{M}_2$.

步骤 6. For j=1 to length (I_{det})

步骤 7. 将 I_{det} 中的每一帧图像输入到手势目标分割模型 $\mathcal{M}_2$ 中, 输出提取手势目标后的图像序列 $I_{\text{seg}} = \{I_1^{'}, I_2^{'}, \cdots, I_n^{'}\}$.

步骤 8. End For

步骤 9. 对分割后的图像进行预处理, 并对图像序列采用最近邻插值方法进行标准化操作, 获得预处理后的图像序列列 $I_{\text{pre}} = \{I_1^{''}, I_2^{''}, \cdots, I_n^{''}\}$.

步骤 10. 依据第 5.2 节的训练策略, 获得用于度量支撑样本和查询样本相似性的 3D 关系神经网络模型 $\mathcal{M}_3$.

步骤 11. 将支撑集 $S^{'}$ 和 I_{pre} 输入到模型 $\mathcal{M}_3$ 中, 并输出样本间的关系分数. 其中, 最高分数对应的索引即为预测的手势类别 y.

步骤 12. 输出动态手势的预测类别 y.

6　综合测试与性能评价

本节利用 DS325 采集的第一视角下手势数据集对本文提出的 OSLHGR 算法性能进行实验验证. 首先, 对用于评估算法性能的手势数据集进行简要介绍, 包括采集环境设置和手势种类. 然后, 对实验方案和网络参数的设置进行说明. 最后, 对实验结果进行综合分析并对算法性能进行评估.

6.1　第一视角手势人机交互的实验测试平台

本文所有实验均使用 Python 作为开发语言, 实验硬件平台是由 Nvidia GTX 1080 GPU 为手势目标的检测和分割模型提供加速运算, 而动态手势分类网络使用 Nvidia Titan Xp 显卡来加速网络模型的训练. 第一视角下手势数据的采集和算法测试是使用 DS325 深度相机完成的. 此外, 我们还基于 TensorFlow 1.3 的 Keras 2.1 和 PyTorch 0.4 的深度学习框架进行深度神经网络模型的开发和应用, 并在 Ubuntu 14.04 上对模型进行深度训练和测试.

6.2　测试数据集的构建

为了评估本文提出的第一视角下基于 3D 卷积神经网络 OSLHGR 算法的性能, 我们利用搭建的手势数据采集实验平台进行了大规模的数据采集工

作, 数据采集环境如图 16(a) 所示. 手势数据采集平台搭建和数据采集过程如下: 1) 首先基于 SoftKinetic DS325 (图 16(b)) 深度相机进行二次开发, 实现对捕获大小为 320 像素 × 240 像素的深度图像以 30 帧/s 的速率进行本地存储; 2) 将深度相机固定在安全帽的正前方, 并穿戴在数据采集者的头部, 同时对深度相机的角度进行微调; 3) 启动应用程序, 受试者在观察实时显示手势电脑桌面的同时, 使用单手或双手进入深度相机的感受野内表演预定义的手势动作, 执行完单个动态手势后双手远离相机感受野区域, 并准备表演第二个手势动作. 如此循环, 直至完成 10 类测试数据的采集工作, 并关闭应用程序. 实际采集的深度图像如图 16(c) 所示.

(a) 手势样本采集实验示例
(a) Experimental example of gesture sample collection

(b) 深度相机 SoftKinetic DS325
(b) SoftKinetic depth sense 325 camera

(c) 采集的深度图像
(c) Depth image collected from the egocentric viewpoint

图 16 数据采集实验平台
Fig. 16 Experimental platform for data collection

不同于文献 [42] 中以纯净的桌面作为表演手势的背景, 本文针对更加实用的应用场景探索基于一次性学习进行手势识别的高效算法. 为此, 在图 16 实验平台的基础上, 采集了 10 类共 500 个第一视角下连续的手势动作作为评估本文算法性能的数据集. 图 17 展示了每一类手势动作示意图. 这些手势的种类和文献 [42] 中选择用于测试算法性能的 10 种手势类别相同, 区别在于数据采集时的环境背景不同. 从图中可以看出, 本文采集的手势数据背景较为复杂, 这会对单样本手势识别算法的性能产生不利的影响. 此外, 按照第 5.2 节中对训练 3D 关系神经网络所使用数据的划分方式, 本文以文献 [42] 中使用的 19 类共 1 995 个手势样本作为训练集, 这 19 种手势的类别和本文采集的手势类别无相互重叠的类.

6.3 测试方案与条件设置

为了对本文提出的第一视角下 OSLHGR 算法的分类性能进行综合分析, 实验方案设计如下. 在

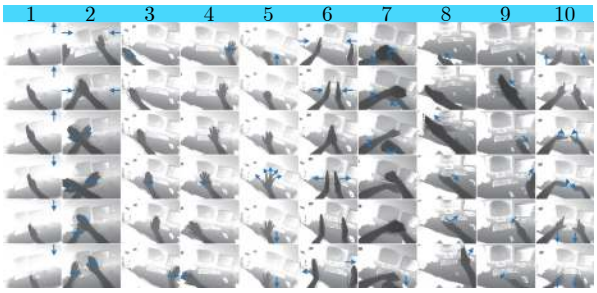


图 17 10 种用于验证 OSLHGR 算法性能的动态手势数据集. 每一列从上向下表示手势核心阶段从起始到结束的变化过程. 图中箭头用于描述动态手势运动的方向

Fig. 17 Ten dynamic gesture datasets to verify the classification performance of OSLHGR algorithm. From top to bottom, each column represents the change process from the beginning to the end of the core phase of gestures. The arrows are used to describe the motion direction of dynamic gestures

使用本文提出的目标检测网络判定手势目标出现在人机交互区域之后, 利用文献 [42−43] 中提出的两阶段算法和本文提出的基于 3D 关系神经网络的单阶段分类算法对第一视角下采集的原始图像序列以及手势目标分割处理后的图像序列分别进行基于一次性学习的动态手势分类实验, 并对比三种算法在手势目标分割前后 OSLHGR 分类的准确率, 验证在复杂背景下基于改进 U-Net 网络模型的手势目标分割与提取对单样本手势识别性能提升的有效性. 同时对文献 [42−43] 和本文算法的分类结果进行对比, 以验证本文提出的算法在模型复杂度、分类准确率和实时性方面的优势. 所有实验输入原始图像的大小均为 320 像素 × 240 像素. 为了进行公平比较, 改进的 SSD 检测网络将原图调整为 300 像素 × 300 像素, 手势目标分割网络输入为 224 像素 × 224 像素. 此外, 由于计算机显存的限制, 3D 关系神经网络将原图调整为 112 像素 × 112 像素, 并将连续 16 帧图像组成一个视频片段用于训练网络模型. 在所有基于 3D 关系神经网络的 OSLHGR 实验中, 初始学习率均设为 10^{-3}, 每迭代 5×10^4 个 episodes 学习率衰减为原来十分之一, 共迭代 1×10^5 次.

6.4 测试结果与性能评价

本节使用第 6.2 节采集的第一视角下动态手势数据集来验证本文提出的 OSLHGR 算法的性能. 首先, 基于第 4 节提出的轻量级 U-Net 网络对复杂背景下的手势目标区域进行提取. 在此基础上, 使用不同的 OSLHGR 算法对预处理后的动态手势进行分类识别, 测试结果如表 4 所示. 同时, 表 4 中还

表 4　OSLHGR 算法的分类结果和模型性能对比

Table 4　Comparison of classification results and model performance of OSLHGR algorithms

分类算法	图像类型	5-way 1-shot准确率 (%)	10-way 1-shot准确率 (%)	参数量 (M)	测试时间 (ms/f)
C3D + Softmax[42]	原始图像	86.40	82.70	28.69	7.88
	手势目标分割图像	91.50	84.95		
Lightweight I3D + NN[43]	原始图像	80.70	74.64	2.94	2.85
	手势目标分割图像	96.16	94.24		
本文算法	原始图像	89.44	73.52	2.08	1.04
	手势目标分割图像	94.64	87.20		

给出了未经分割处理的手势分类结果. 通过对比可以得知, 手势目标的精确分割可以大幅降低复杂背景对分类结果的影响, 提升分类准确率, 这对于只有单样本的分类任务而言是至关重要的. 此外, 为了进一步说明本文方法在分类准确率和实时性方面的性能优势, 按照本文使用的测试策略对文献 [42−43] 中的测试方法进行了修改. 表 4 中分别给出了在 5-way 1-shot 和 10-way 1-shot 下的分类结果. 通过与文献 [42] 的分类结果对比可以看出, 本文方法在手势目标分割后数据集上的分类结果明显优于后者, 而在原始图像上 10-way 1-shot 却不及后者. 这主要由于文献 [42] 使用了连续微调的训练机制, 每次都从新的手势类中随机选择单个样本微调网络的分类层, 故在元训练集和元测试集背景不同的情况下, 表现出较好的分类性能. 此外, 该方法的网络参数量、时间开销和内存占用远超本文提出的分类算法. 与文献 [43] 的分类结果对比发现, 本文算法在手势目标分割提取后的数据样本上 10-way 1-shot 分类准确率要低于前者. 而在原始图像上 10-way 1-shot的分类准确率同文献 [43] 具有相当的分类性能, 且 5-way 1-shot 的分类准确率 89.44% 远高于 80.70%. 这是由于文献 [43] 采用两阶段的分类策略, 在训练网络的基础类数据和验证模型性能的测试数据背景不一致时, 无法对预训练模型参数进行调节, 导致分类性能大幅降低. 而本文采用了单级式基于 episode 的训练策略, 可有效地降低因数据差异对分类性能产生的影响. 通过对三种分类算法的对比可知, 本文提出的算法在保持较低参数量和较高实时性的同时, 在分类准确率上也保持在较为满意的水平, 本文算法的有效性得到了充分的验证.

7　总结与展望

本文提出了一种基于深度神经网络的级联组合进行 OSLHGR 的分类算法, 以实现第一视角下手势动作的快速和精确分类, 提升智能人机交互中的用户体验. 在该算法中, 为了满足在便携式移动系统中的应用和实现手势目标快速精准检测的需求, 运用 MobilenetV2 对端到端 SSD 目标检测模型进行轻量化设计, 并将编−解码架构、感受野区块和门控单元加入到检测网络, 在 Pascal VOC 2012数据集和 SoftKinetic DS325 采集的手势目标检测数据集上分别达到 73.6% 和 96.3% 的均值平均精度, 实现了轻量级模型检测性能的大幅提升. 进而, 为了有效降低复杂背景的干扰, 提升 OSLHGR 算法的性能, 本文提出的轻量级 U-Net 网络在手势目标分割数据集上的交并比为 94.53% 且计算复杂度和处理速度等性能指标均表现优异. 在精确分割手势目标的基础上, 本文提出的 3D 关系深度神经网络实现了对第一视角下动态手势的有效分类, 取得了 94.64% 的 5-way 1-shot 识别准确率, 这为复杂应用环境下便捷式智能人机交互提供了可靠的技术保障.

本文提出的基于深度神经网络级联组合实现第一视角下一次性学习手势识别的算法还可推广到车载影音控制系统、垃圾分类的体感游戏等智能人机交互场景. 此外, 针对交互过程中在完成某个动作后手需离开相机感受野这一限制, 在后续工作中, 我们将针对复杂场景下连续动作的 OSLHGR 展开研究, 以降低手势表达的约束条件, 实现更加便捷自然的智能人机交互.

References

1　Betancourt A, López M M, Regazzoni C S, Rauterberg M. A sequential classifier for hand detection in the framework of egocentric vision. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, Ohio, USA: IEEE, 2014. 600−605

2　Thalmann D, Liang H, Yuan J. First-person palm pose tracking and gesture recognition in augmented reality. *Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2015, **598**: 3−15

3　Serra G, Camurri M, Baraldi L, Benedetti M, Cucchiara R. Hand segmentation for gesture recognition in EGO-vision. In: Proceedings of the 3rd ACM International Workshop on Interactive Multimedia on Mobile and Portable Devices. Barcelona, Spain: ACM, 2013. 31−36

4　Cao C Q, Zhang Y F, Wu Y, Lu H Q, Cheng J. Egocentric gesture recognition using recurrent 3D convolutional neural net-

works with spatiotemporal transformer modules. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 3763−3771

5 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. SSD: Single shot multiBox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 21−37

6 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 2015 International Conference on Medical Image Computing and Computer-Assisted Intervention. Munich, Germany: Springer, 2015. 234−241

7 Sung F, Yang Y X, Zhang L, Xiang T, Torr P H S, Hospedales T M. Learning to compare relation network for few-shot learning. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 1199−1208

8 Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 4510−4520

9 Laibacher T, Weyde T, Jalali S. M2U-Net: Effective and efficient retinal vessel segmentation for resource-constrained environments. arXiv preprint, arXiv: 1811.07738, 2018.

10 Ibtehaz N, Rahman M S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 2020, **121**: 74−87

11 Peng Yu-Qing, Zhao Xiao-Song, Tao Hui-Fang, Liu Xian-Zi, Li Tie-Jun. Hand gesture recognition against complex background based on deep learning. *Robot*, 2019, **41**(4): 534−542
(彭玉青, 赵晓松, 陶慧芳, 刘宪姿, 李铁军. 复杂背景下基于深度学习的手势识别. 机器人, 2019, **41**(4): 534−542)

12 Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 779−788

13 Yip H M, Navarro-Alarcon D, Liu Y. Development of an eye-gaze controlled interface for surgical manipulators using eye-tracking glasses. In: Proceedings of the 2016 IEEE International Conference on Robotics and Biomimetics. Qingdao, China: IEEE, 2016. 1900−1905

14 Wanluk N, Visitsattapongse S, Juhong A, Pintavirooj C. Smart wheelchair based on eye tracking. In: Proceedings of the 9th Biomedical Engineering International Conference. Luang Prabang, Laos: IEEE, 2016. 1−4

15 Yang Guan-Ci, Yang Jing, Su Zhi-Dong, Chen Zhan-Jie. An improved YOLO feature extraction algorithm and its application to privacy situation detection of social robots. *Acta Automatica Sinica*, 2018, **44**(12): 2238−2249
(杨观赐, 杨静, 苏志东, 陈占杰. 改进的 YOLO 特征提取算法及其在服务机器人隐私情境检测中的应用. 自动化学报, 2018, **44**(12): 2238−2249)

16 Li Chang-Ling, Li Wei-Hua. A multimodal interaction model for battlefield. *Fire Control and Command Control*, 2014, **39**(11): 110−114
(李昌岭, 李伟华. 面向战场的多通道人机交互模型. 火力与指挥控制, 2014, **39**(11): 110−114)

17 Zhang Y F, Cao C Q, Cheng J, Lu H Q. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, 2018, **20**(5): 1038−1050

18 Hegde S, Perla R, Hebbalaguppe R, Hassan E. GestAR: Real time gesture interaction for AR with egocentric view. In: Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality. Merida, Mexico: IEEE, 2016. 262−267

19 Bambach S, Bambach S, Crandall D J, Yu C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1949−1957

20 Pandey R, White M, Pidlypenskyi P, Wang X, Kaeser-Chen C. Real-time egocentric gesture recognition on mobile head mounted displays. In: Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA, USA: IEEE, 2017. 1−4

21 Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyamd T, Andreetto M, Adam H. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint, arXiv: 1704.04861, 2017.

22 Zhang Hui, Wang Kun-Feng, Wang Fei-Yue. Advances and perspectives on applications of deep learning in visual object detection. *Acta Automatica Sinica*, 2017, **43**(8): 1289−1305
(张慧, 王坤峰, 王飞跃. 深度学习在目标视觉检测中的应用进展与展望. 自动化学报, 2017, **43**(8): 1289−1305)

23 Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, Ohio. USA: IEEE, 2014. 580−587

24 Girshick R. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1440−1448

25 Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(6): 1137−1149

26 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 2015 International Conference on Learning Representations. San Diego, USA: ICLR, 2015. 1−14

27 Fu C Y, Liu W, Tyagi A, Berg A C. DSSD: Deconvolutional single shot detector. arXiv preprint, arXiv: 1701.06659, 2017.

28 Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA: IEEE, 2015. 1−9

29 Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection. In: Proceedings of the 2018 European Conference on Computer Vision. Munich, Germany: Springer, 2018. 385−400

30 Shen Z Q, Shi H, Feris R, Cao L L, Yan S C, Liu D, et al. Learning object detectors from scratch with gated recurrent feature pyramids. arXiv preprint, arXiv: 1712.00886, 2017.

31 Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 7132−7141

32 Zhang Xue-Song, Zhuang Yan, Yan Fei, Wang Wei. Status and development of transfer learning based category-level object recognition and detection. *Acta Automatica Sinica*, 2019, **45**(7): 1224−1243
(张雪松, 庄严, 闫飞, 王伟. 基于迁移学习的类别级物体识别与检测研究与进展. 自动化学报, 2019, **45**(7): 1224−1243)

33 Wang T, Chen Y, Zhang M Y, Chen J, Snoussi H. Internal

transfer learning for improving performance in human action recognition for small datasets. *IEEE Access*, 2017, **5**(1): 17627−17633

34 Li Y X, Li J W, Lin W Y, Li J G. Tiny-DSOD: Lightweight object detection for resource-restricted usages. In: Proceedings of the 2018 British Machine Vision Conference. Newcastle, UK: BMVC, 2018. 1−12

35 Wong A, Shafiee M J, Li F, Chwyl B. Tiny SSD: A tiny single-shot detection deep convolutional neural network for real-time embedded object detection. In: Proceedings of 15th Conference on Computer and Robot Vision. Toronto, Canada: IEEE, 2018. 95−101

36 Liau H, Yamini N, Wong Y L. Fire SSD: Wide fire modules based single shot detector on edge device. arXiv preprint, arXiv: 1806.05363, 2018.

37 Wang R J, Li X, Ling C X. Pelee: A real-time object detection system on mobile devices. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: IEEE, 2018. 1967−1976

38 He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 1026−1034

39 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **39**(4): 640−651

40 Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, **39**(12): 2481−2495

41 He K M, Gkioxari G, DollárP, Girshick R. Mask R-CNN. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2961−2969

42 Lu Z, Qin S Y, Li X J, Li L W, Zhang D H. One-shot learning hand gesture recognition based on modified 3D convolutional neural networks. *Machine Vision and Application*, 2019, **30**(7–8): 1157−1180

43 Lu Z, Qin S Y, Li L W, Zhang D H, Xu K H, Hu Z Y. One-shot learning hand gesture recognition based on lightweight 3D convolutional neural networks for portable applications on mobile systems. *IEEE Access*, 2019, **7**: 131732−131748

鹿 智 北京航空航天大学自动化科学与电气工程学院博士研究生. 2016 年获得北京信息控制研究所计算机科学与技术硕士学位. 主要研究方向为机器视觉和模式识别.
E-mail: by1603117@buaa.edu.cn
(**LU Zhi** Ph.D. candidate at the School of Automation Science and Electrical Engineering, Beihang University. He received his master degree in computer science and technology from Beijing Institute of Information Control in 2016. His research interest covers machine vision and pattern recognition.)

秦世引 北京航空航天大学自动化科学与电气工程学院教授和东莞理工学院电子工程与智能化学院教授. 1990 年获得浙江大学工业控制工程与智能自动化博士学位. 主要研究方向为模式识别与机器学习, 图像处理与计算机视觉, 人工智能及其应用. 本文通信作者. E-mail: qsy@buaa.edu.cn
(**QIN Shi-Yin** Professor at the School of Automation Science and Electrical Engineering, Beihang University, also at the School of Electrical Engineering and Intelligentization, Dongguan University of Technology. He received his Ph.D. degree in industrial control engineering and intelligent automation from Zhejiang University in 1990. His research interest covers pattern recognition and machine learning, image processing and computer vision, and artificial intelligence and its applications. Corresponding author of this paper.)

李连伟 北京航空航天大学自动化科学与电气工程学院博士研究生. 2017 年获得山东大学控制科学与工程学院学士学位. 主要研究方向为深度学习和计算机视觉.
E-mail: llw2017@buaa.edu.cn
(**LI Lian-Wei** Ph.D. candidate at the School of Automation Science and Electrical Engineering, Beihang University. He received his bachelor degree from the School of Control Science and Engineering, Shandong University in 2017. His research interest covers deep learning and computer vision.)

张鼎豪 北京航空航天大学电子信息工程学院硕士研究生. 2018 年获得北京航空航天大学自动化学士学位. 主要研究方向为计算机视觉和模式识别. E-mail: hbhszdh@buaa.edu.cn
(**ZHANG Ding-Hao** Master student at the School of Electronic Information Engineering, Beihang University. He received his bachelor degree in automation from Beihang University in 2018. His research interest covers computer vision and pattern recognition.)