软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(10):2986−3004 [doi: 10.13328/j.cnki.jos.005785]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 无菌条件非接触式多通道自然交互手术环境*

陶建华[1], 杨明浩[1], 王志良[2], 班晓娟[2], 解 仑[2], 汪云海[3], 曾 琼[3], 王 飞[4], 王红迁[4], 刘 斌[1], 韩志帅[2], 潘 航[2], 陈文拯[3]

[1](模式识别国家重点实验室(中国科学院 自动化研究所),北京 100190)

[2](北京科技大学 计算机与通信工程学院,北京 100083)

[3](山东大学 计算机科学与技术学院,山东 青岛 266237)

[4](陆军军医大学 重庆西南医院 信息科,重庆 400038)

通讯作者: E-mail: mhyang@nlpr.ia.ac.cn

**摘 要**: 无菌和非接触环境是医疗手术室的基本要求,这使得计算机操作室和手术室需要在物理上隔离.同时,因为手术进行中,主治医生如果需要查看病灶图像,通常授意护士或者手术助理到计算机操作室操作病灶图像,由于手术室和计算机操作室间的隔离,以及主治医生和助理间可能存在的意图理解不准确,容易导致护士或者手术助理在手术室和计算机操作室往返多次,这增加了患者手术时间延长、失血增多、脏器暴露时间长等风险,尽量减少手术中定位到病灶图像的时间对于医生和病人都很重要.针对上述需求,借助遮挡环境下的深度图像人体骨架提取、手势跟踪与理解、手术室环境远场语音识别,多模态信息处理与融合技术,构建无菌条件下的非接触式多通道自然交互手术环境.该环境使得主治医生在手术中可通过语音命令、手势及上述交互方式相结合的方式快速定位到需要观察的病灶成像.在接近真实环境的实验环境中,建立的无菌条件的非接触式多通道自然交互手术环境在保证精度的情况下,可显著缩短病灶图像定位时间.无菌环境智能交互医疗手术室为未来下一代高效的手术提供了技术与方法验证.

**关键词**: 手术室;多模态信息融合;意图理解

**中图法分类号**: TP391

中文引用格式: 陶建华,杨明浩,王志良,班晓娟,解仑,汪云海,曾琼,王飞,王红迁,刘斌,韩志帅,潘航,陈文拯.无菌条件非接触式多通道自然交互手术环境.软件学报,2019,30(10):2986−3004. http://www.jos.org.cn/1000-9825/5785.htm

英文引用格式: Tao JH, Yang MH, Wang ZL, Ban XJ, Jie L, Wang YH, Zeng Q, Wang F, Wang HQ, Liu B, Han ZS, Pan H, Chen WC. Non contact multi-channel natural interactive surgical environment under sterile conditions. Ruan Jian Xue Bao/Journal of Software, 2019,30(10):2986−3004 (in Chinese). http://www.jos.org.cn/1000-9825/5785.htm

## Non Contact Multi-channel Natural Interactive Surgical Environment under Sterile Conditions

TAO Jian-Hua[1], YANG Ming-Hao[1], WANG Zhi-Liang[2], BAN Xiao-Juan[2], XIE Lun[2], WANG Yun-Hai[3], ZENG Qiong[3], WANG Fei[4], WANG Hong-Qian[4], LIU Bin[1], HAN Zhi-Shuai[2], PAN Hang[2], CHEN Wen-Zheng[3]

[1](National Laboratory of Pattern Recognition (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

[2](School of Computer & Communication Engineering, University of Science & Technology Beijing, Beijing 100083, China)

[3](School of Computer Science and Technology, Shandong University, Qingdao 266237, China)

[4](Information Department, Southwest Hospital, Army Medical University, Chongqing 400038, China)

*基金项目: 国家重点研发计划(2016YFB1001404); 国家自然科学基金(61873269, 61831022, 61425017, 61332017)

Foundation item: National Key Research & Development Plan of China (2016YFB1001404); National Natural Science Foundation of China (61873269, 61831022, 61425017, 61332017)

本文由“自然人机交互新进展”专题特约编辑田丰、喻纯推荐.

收稿时间: 2018-08-18; 修改时间: 2018-11-01; 采用时间: 2018-12-25; jos 在线出版时间: 2019-04-29

CNKI 网络优先出版: 2019-04-30 10:09:23, http://kns.cnki.net/kcms/detail/11.2560.TP.20190430.1009.011.html

© 中国科学院软件研究所 http://www.jos.org.cn

**Abstract**: Sterile and non-contact environment is the basic requirement of medical operating room, which makes the computer room and operating room need to be physically isolated. At the same time, if the attending doctor needs to look at the image of the lesion during the operation, he usually instructs the nurse or the assistant to operate the image of the lesion in the computer operating room, because of the isolation between the operating room and the computer room, and because the intention between the attending doctor and the assistant may not be understood accurately, it is easy to lead the nurse or surgical assistant in the operating room and computer room to and fro many times, which increases the risk of prolonged operation time, increased blood loss, and organ exposure time, minimizing the time to locate the lesion image in the operation is important for doctors and patients. To meet the above requirements, a non-contact multi-channel natural interactive surgical environment under aseptic conditions is constructed by means of human skeleton extraction, gesture tracking and understanding, far-field speech recognition in operating room environment, multi-modal information processing, and fusion technology. This environment allows the attending physician to quickly locate the lesion to be observed during surgery by combining voice commands, gestures, and the above interaction. In the experimental environment close to the real environment, the non-contact multi-channel natural interactive surgical environment established in this study can significantly reduce the localization time of the lesion image under the condition of ensuring the accuracy. Intelligent interactive operating room in aseptic environment provides technical and methodological validation for the next generation of efficient surgery.

**Key words**: operation room; multimodal information fusion; intention understanding

## 1 介 绍

医疗卫生信息化建设进程使得外科手术向微创及精准化发展,同时,对手术的安全性和舒适性提出了更高的要求.下一代手术室功能不但要满足手术需求,还要体现现代化医院的设施水平、医疗水平和管理水平,同时还需要将洁净化、数字化和人性化融为一体.现代化手术室建设涉及室内环境整合及控制、手术视音频信号采集分配管理、手术及相关设备控制、医疗影像诊断资料的采集传输存储、医院信息系统集成及远程交互等多方面的内容,是涵盖医院多科室联合的综合系统工程[1],历史上的国内外手术室的发展历程可大概分为 4 个阶段.第 1 阶段:传统手术室,一般是仅能对病人实施局部麻醉的小手术,不需要太多的仪器设备的接入,信息基本都是人工采集记录.第 2 阶段:现代手术室,一般都可实现对病人的复苏照顾,部分信息化设备已经逐步加以应用,可做的手术越来越多且相对复杂.第 3 阶段:数字化手术室,起源于 20 世纪 90 年代,一般就是在目前洁净手术室的基础上,综合应用各种信息化设备和软件技术,实现通过设备来采集数据、监控病人状态,部分远程示教,实现部分信息的共享[2].第 4 阶段:智能数字化手术室,实现手术室内部的非接触式手术识别,远程示教、手术全过程信息的管理,达到手术医生可不离开手术台即可精准、实时地获取病人的一切相关信息,医生可通过信息实时、动态地掌握每一个手术详细的细节,病人家属也可相对更加详细地获得手术进度,下一步实现人与机器更加紧密的结合,共同完成手术[3].

在上述建设条件中,无菌和非接触环境是构建医疗手术室的基本要求,这样的要求使得计算机操作室和手术室通常在大多数情况下距离很近,但会在物理上隔离开.在手术中,主治医生通常需要查看病灶图像,如患者手术前图像细节,如血管、神经、周围临近器官的空间位置等.结构越复杂的手术,主治医生会在查看病灶上花的时间越多,以乳腺癌肿瘤手术切除为例,目前每台手术在上述环节总体需要耗时 20 分钟~1 小时不等,患者手术时间延长会导致术中失血增多,脏器暴露时间长会增加感染,不利于患者的术后恢复,另外还会增加手术后并发症的风险[4].传统手术中,主治手术医师通常是通过授意护士或者手术助理到计算机操作室操作病灶图像.因为手术室和计算机操作室间的距离,以及手术室主治医生和助理间不熟悉程度可能存在的意图理解错误风险,容易导致患者手术时间延长、失血增多、脏器暴露时间长等风险,因此,尽量减少定位到病灶图像的时间对于医生和病人都很重要.

近年来,随着人工智能技术的发展,如语音识别技术[5,6]、姿态跟踪与理解[7−9]、手势理解[10−13]、多模态信息融合技术等[14−19],这些技术与方法使得用户可以通过非接触式的方式与计算机交互,为建立新型无菌条件的非接触式自然交互手术室提供了方法与技术上的可能.然而,在手术室环境中利用上述技术仍然存在许多挑战:(1) 手术室环境要求医生的穿戴尽量简洁;(2) 手术台面以及护士及助理使得主治医生的姿态处于遮挡环境,为准确地进行姿态跟踪带来了挑战;(3) 因为手术环境血液污染等,为手势跟踪及手势的准确理解带了困难.

© 中国科学院软件研究所 http://www.jos.org.cn

尽管最新的人工智能技术取得了很大进展,但要很好地将这些技术应用到新型无菌条件的非接触式自然交互手术室仍然存在诸多困难.有研究认为,恰当的多通道融合的交互方式在表达效率和完整性上都要优于单一模式[20],因此,如何在上述交互通道上,通过多通道信息融合的模式建立非接触交互的智能手术室,提高手术环境下计算机对主治医师的交互意图理解,实现无菌状态下手术器械及材料的准确传递与自然、高效率的病灶图像查阅,在减少传统手术室的过多人环节的同时合理缩短手术时间,提高各个环节的效率和质量,最大程度地消除无菌手术人员和非手术人员的交流障碍,实现手术室内部的非接触式手术识别,达到手术医生可不离开手术台即可精准、实时地获取病人相关信息,是下一代智能手术室的重要需求[1−4].针对上述需求,本文通过融合遮挡环境下的深度图像人体骨架提取、手势跟踪与理解、手术室环境远场语音识别,多模态信息处理与融合技术,构建了无菌条件下的非接触式多通道自然交互手术环境,使得主治医生在手术中可通过语音命令、手势及上述几种交互方式相结合快速定位到需要观察的病灶成像.在接近实际的实验环境中,本文建立的无菌条件的非接触式多通道自然交互手术环境在保证精度的情况下,可显著缩短病灶图像的定位时间.

本文第 2 节介绍相关工作.第 3 节介绍无菌条件下的非接触式多通道自然交互手术环境的技术总体框架.第 4 节和第 5 节分别介绍面向无菌自然交互手术室各单一通道技术、多通道信息融合相关理论与方法.第 6 节介绍相关实验、结果及分析.第 7 节给出本文的总结及展望.

## 2 相关工作

无菌条件下的非接触式多通道自然交互手术环境主要基于遮挡环境下的深度图像人体骨架提取、交互手势理解、手术室环境远场语音识别,多模态信息处理与融合技术等技术构建.本节介绍相关技术工作,并分析目前相关技术用于自然交互手术环境所存在的挑战.

### 2.1 遮挡条件下人体骨架提取

人体骨架提取算法主要依赖于光学相机所采集的图像或视频信息,利用图像或视频特征算子[21,22]获取人体二维骨架.然而,此类算法所提取的骨架精度受限于特征算子应用假设的约束,且不可避免具有二维信息场所具备的空间局限性,无法表达三维相关的信息(比如遮挡),因此,难以满足实际应用需求[23].近年来,随着三维扫描技术的日益成熟,愈来愈多的算法利用三维深度信息提取三维人体骨架,通过融合激光扫描仪、深度相机等设备采集的三维信息,采用几何处理的办法提取人体三维骨架.目前,人体骨架提取的研究正逐步由静态的简单结构化场景向动态的非结构化群体遮挡复杂场景转化,由二维骨架提取发展为三维骨架提取.然而,这类方法存在的问题在于:激光扫描仪不仅造价过高,而且所获取的点云存在较多噪声,难以与图像信息匹配,不适合于复杂场景;深度相机由于其硬件的限制,仅能够获取一定范围内的深度信息,且无法精细化处理远距离场景下手势等细粒度应用[24,25].

为了从无标记运动采集数据中提取精确的三维人体骨架,深度学习理论与技术的发展为解决这一问题提供了重要思路[26−28].Belagiannis 等人[29]基于手术室场景中布置的多个光学相机提取人体三维骨架,该方法首先构造手术室场景下的二维人体姿态库,利用卷积神经网络,根据目标检测所获取的人体提取相应二维骨架,然后基于条件随机场(conditional random field)以结构化支持向量机(structure SVM)及将不同视角获取的二维姿态对应到三维人体骨架.然而,该方法依赖于目标检测算法,其分阶段式的三维姿态估计(先估计二维姿态,再转换成三维姿态)易造成不同阶段的累积误差.Kadkhodamohammadi 等人[30]探索了手术室场景下基于单视角 RGB-D 提取人体三维骨架的算法,该方法拓展了传统骨架提取图结构(pictorial structure)[31]框架,利用 RGB 信息构建表面模型以及三维深度约束构建形变模型,并提出了差分直方图作为深度图像的特征.同年,Kadkhodamohammadi 等人[32]通过结合卷积神经网络提取特征表达、基于随机森林的姿态及位置先验估计以及多视角优化,更进一步地将该算法拓展应用至基于多视角 RGB-D 的手术室场景.然而,此类算法受限于深度相机的数据精度以及深度学习所需要的标记数据.相比深度相机采样尺度限制,二维图像数据具备高精度、高清晰度等特性,能够从多尺度适应手术室的复杂环境.尽管相关算法取得了一定进展,但要准确地从无标记运动采集数据中提取精确的三维人体骨架,依然是一个极具挑战性的问题.

© 中国科学院软件研究所 http://www.jos.org.cn

### 2.2 交互手势理解

手势交互首先需要识别人体手势,常用的手势识别算法可以是非模板匹配算法,也可以基于模板匹配.模板匹配算法不易混淆手势,并且在训练数据很少的情况下也能够达到较高的准确率.Ruan 等人[33]从动态时间规整(dynamic time warping,简称 DTW)算法的约束条件出发提出了放宽端点对齐和全局路径限制的方案,针对 DTW 算法,其速度和准确率都有较大的提升.Chao 等人[34]在传统 DTW 算法的基础上根据每个骨骼节点对手势贡献的不同分别推算了加权距离,提升了识别准确率,在复杂背景和光照方面有很好的鲁棒性.Wu 等人基于 DTW 和 *K*-means 进行人体动作匹配和评估,完成了病人康复训练系统.Pan 等人[35]利用改进的 DTW 算法实现了在线人体动作识别,Hiyadi 等人[36]使用自适应滑动窗口与 DTW 结合的方式,能够识别出混合手势动作中的所有简单手势.由此可见,DTW 算法无需过多样本进行训练,只需要确定好手势模板,便能够达到较好的性能,在手势识别结果的混淆程度上也低于一般的非模板匹配算法,但是它无法识别连续的重复手势,这将导致用户无法对同一张图片进行连续的放大、移动等操作.

在非模板匹配算法方面,Zhang 等人[37]通过支持向量机(support vector machine,简称 SVM)对 Kinect 产生的骨骼数据进行分类,目前已经实现了 22 种姿势的识别.Chen 等人[38]使用 SVM 实现了实时识别人手画出的 0~9 等数字以及 26 个英文字母,Zhang 等人[39]使用隐马尔可夫模型(hidden Markov model,简称 HMM)[40]实现了手势轨迹的识别,Song 等人[41]使用高斯混合模型(Gaussian mixture model,简称 GMM)和 HMM 完成了全身姿势的实时识别,Wang 等人[42]使用卷积神经网络(convolutional neural network,简称 CNN)实现了大规模的连续手势识别,Li 等人[43]使用主成分分析法(principal components analysis,简称 PCA)结合 CNN 实现了对中国人表达数字的相关手势的识别,Chavan 等人[44]使用“随机森林(random forest,简称 RF)”对印度的手语手势进行分类,在连续手势中能够提取表达手语意义的片段并显示结果.这些非模板匹配算法在进行手势识别之前都需要进行训练,在识别过程中容易对连续手势动作产生混淆,若手势样本过少,将对非模板匹配相关算法的性能产生很大影响.

### 2.3 远场语音识别

远场环境下录制的语音会面临非平稳噪声和高混响的干扰,从而导致话音质量的下降,直接影响到语音识别的性能.在算法方面,基于麦克风阵列的波束形成技术已得到很多年的发展,需要解决的核心问题是协方差矩阵的计算和导向矢量的估计,比较经典的方法包括加权延时求和法[45]、最小方差失真响应法[46]、广义旁瓣滤波法[47]、多通道维纳滤波法[48]等.随着深度学习在语音领域的广泛应用,相继有一些基于深层神经网络的多通道语音增强算法[49,50]被提了出来,以实现对非平稳噪声和非目标方向干扰源的抑制,但上述方法大多受限于硬件结构,其性能仍有较大的提升空间;远场语音处理中的另一难点是混响抑制,不同房间对应不同的混响函数,仿真生成的混响数据和真实混响数据存在较大的差异,使得混响比噪声更难处理,主流的混响抑制方法包括谱减法[51]、加权预测误差法[52]、深层神经网络法[53]等,上述方法虽然能够抑制混响干扰,但当噪声和混响同时存在时,算法性能明显下降.通过前端和后端联合优化建模是提高远场语音识别性能的有效途径[54,55].前端的信号处理技术一般只用到当前状态下的语音的信号信息,这些信息的利用主要依靠对声学物理规律的把握,并基于一定的假设,而机器学习的方法能够利用很多的训练集里学到的信息来建模,但是它一般不是基于物理原理的,对当前帧信息的使用比较弱.所以,把这两种方法比较好地融合在一起是目前很多研究机构发力的一个方向.一种典型的方式是把前端的信号处理与后端的语音识别引擎进行更好的联合优化[56],前端信号处理有可能丢失信息且不可在后端恢复,而分别优化的策略可能对于前端来说是最优的,但对于整个系统未必是最优选项.因此需要一种有效的建模方法,以使前端可以有效提升信号质量但同时比较少地丢失信息,而把一些剩余的噪声留给更强大的后端来处理,从而提升整体性能[57,58].

### 2.4 多模态信息处理与融合

多通道信息融合方法按照发生的时间顺序,可以分为前期融合和后期融合;按照信息融合的层次来分,融合可以分别发生在数据(特征)层、模型层及决策层;如果按照处理方法来分,可分为基于规则的融合,或者基于统计(机器学习方法)的融合.也有文献根据多通道信息的相关性,把它们的关系分为信息互补、信息互斥、信息冗余

© 中国科学院软件研究所 http://www.jos.org.cn

这样几个特点,然后根据其信息特点分别加以融合.

数据层、特征层、决策层的融合方法偏重于模型的设计,同时,在多模态信息融合的计算方法中大都通过采用基于统计和机器学习的方法进行模型的构建,如贝叶斯决策模型、神经网络模型、图模型等等.贝叶斯决策模型的特点在于其能够根据不完全情报,对部分未知的状态采用主观概率估计,然后用贝叶斯公式对发生概率进行修正,最后利用期望值和修正概率做出最优决策[59].在多种通道信号联合分布概率部分已知的情况下,贝叶斯决策模型可以根据历史经验反演得到某些缺失的信号,从而得到整个多通道信号融合整体最优评估.传统的神经网络模型在非线性函数拟合方面表现出很好的性能,并在单一通道的信息处理上,深度神经网络模型取得了很好的效果,因此,很多研究者希望综合不同的神经网络模型,如 LSTM、CNN、RNN 结构,构建面向多通道信息融合的大规模深度神经网络模型,力图在融合阶段无差别地处理多通道信息.图模型将概率计算和图论结合在一起,提供较好的不确定性计算工具,其构成上的节点以及节点之间的连线,使其在计算变量与周围相连变量的关系上具有一定优势.相对于无向图模型,有向图模型节点之间的连线不仅记忆了数据流向,还记录有学习过程中的状态跳转概率,有向图模型除了可以用于不确定性计算外,还可用于面向时序问题的决策推理,如基于动态贝叶斯模型模仿产生人类对文字的书写过程[60]等.除了以上多通道信息融合计算模型外,还有很多其他模型也用于多通道信息融合,如多层支持向量机、决策回归树、随机森林等方法.

## 3 研究框架

整个算法框架如图 1 所示,输入部分主要为包含姿态、手势、语音的 3 个主治医师交互通道信息.姿态模块用于在遮挡条件下准确地提取人体的框架,进而识别出医生的姿态;手势模块用于获取医生的手部动作并识别出特定的手势;语音模块完成基于麦克风阵列的远场语音识别,并转化成指令.3 个输入模块的信息进行多通道的信息融合,实现医生的意图分类和理解,将分类结果通过交互界面反馈给医生,下面分别介绍各单一模态信息处理技术及信息融合方法.

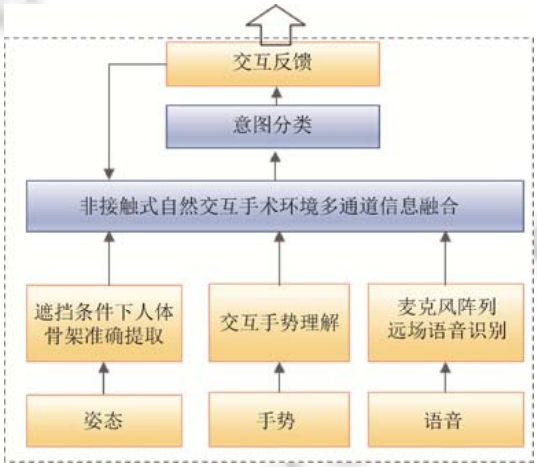



Fig.1 The framwwork of non contact multi-channel natural interactive surgical environment under eterile condition

图 1 无菌条件非接触式多通道自然交互手术环境整体研究框架

## 4 无菌条件下的不同通道信息感知方法

### 4.1 手术室遮挡条件下人体骨架准确提取

手术室场景存在较多的环境干扰(比如非自然灯光、复杂手术设备以及缺乏纹理信息的手术服),且场景中医生、护士、病人等人员彼此之间存在大量复杂的遮挡及自遮挡关系,是动态的非结构化群体复杂场景.因此,

© 中国科学院软件研究所 http://www.jos.org.cn

如何在手术室群体复杂场景下高精度地提取人体骨架是一个极具挑战性的问题.另外,手术室场景具有环境多样、遮挡复杂等特性,除此以外,由于手术情况下应该尽可能地减少“侵入性”设备的使用,手术场景下的数据采集系统通常都是无标记的运动采集系统(比如光学相机、深度相机),给手术场景下提取人体三维骨架提出了更高要求.因此,我们认为手术室场景应该充分利用光学相机采集的二维图像信息.然而,由二维图像估计三维人体骨架是一个病态问题,尽管深度学习为解决该问题提供了有利工具,但却面临着三维姿态训练数据缺失的问题.为此,我们提出了一种全自动的、大规模人体姿势空间采样并生成人体三维姿势训练集合的算法,基于深度学习端对端特性从单张二维图像中全自动地提取三维人体骨架.该算法主要涉及人体三维姿态数据集合成、人体三维姿态回归以及人体三维骨架提取这三大步骤.

(1) 人体三维姿态数据集合成

针对三维人体骨架训练数据极难标注的问题,我们在三维模型集合上大规模地渲染人体图片及相应骨架标签.我们认为,合成数据集中的人体姿态分布应当与真实图像中的人体姿态分布相一致.为了更为完整地覆盖整个人体空间,需要根据已有动作推断自然连续的未知动作.我们发现,自然动作往往与联合变化的人体部位相关(比如胳膊的前臂和后臂),可通过组合人体部位生成新的姿态.因此,我们利用基于运动捕捉设备捕获的姿态以及二维图像中恢复的姿态为样本,学习了一个稀疏的、非参数化的贝叶斯模型[61]以分解人体姿态表达,通过组合人体子关节结构生成新的姿态,从而生成更为丰富的模型表达.由此获取的人体三维姿态利用现有算法(如SCAPE 模型[62])生成三维模型,通过添加不同的纹理贴图能够生成丰富的人体姿态图像.经过人体姿态采样与纹理迁移后,我们能够合成不同姿态、不同纹理的人体模型,通过改变渲染视角、渲染背景灯,能够合成与真实图片高度一致的二维图像.多样化二维图像与人体三维姿态的对应,为基于深度学习的单幅图像三维人体骨架的提取提供了数据基础.

(2) 人体三维姿态回归域迁移网络

为了避免由于真实图像与合成图像的差异所带来的过拟合、最大优化训练性能,我们提出一种域迁移网络回归真实图像中的人体三维姿态,其核心思想在于将渲染图像与真实图像投影到相同特征空间,从而缩小渲染图像与真实图像集之间的分布差异.如图2蓝色虚线部分所示,该域迁移网络主要包括3部分结构:特征提取器、姿态回归器以及域间分类器.特征提取器主要负责提取图像特征,采用了AlexNet[63]的conv1到pool5层作为特征提取网络(这里可用其他卷积神经网络代替).该特征被同时输入至姿态回归器及域间分类器,其中,姿态回归器用于判别三维姿态.域间分类器用于判别高维特征的真实性,以促使特征提取器提取与真实图像一致的特征.

域迁移网络采用对抗网络分阶段训练思想训练模型,输入包含具有三维姿态标签的渲染图像以及没有三维姿态标签的真实图像.训练分为两个阶段:第1个阶段(图2上半部分所示),我们固定特征提取器,输出特定的特征用于训练姿态回归器和域间分类器.姿态回归器用于回归三维姿态,域间分类器用于判别图像类别(即真实图像还是合成图像).第2个阶段(图2下半部分所示),我们固定域间分类器,训练特征提取器和姿态回归器.这里,要求特征提取器输出一种新的特征,该特征能够保持下述约束:(1) 可用于姿态回归器回归三维姿态;(2) 域间分类器能够依据该特征输出(0.5,0.5)的类别判断,该约束的目的在于“迷惑”域间分类器,使其无法判别出图片类别.训练域迁移网络至域间分类器无法判别图像类别,则说明真实图像和合成图像的特征属于同一特征空间,缩小了合成图像与真实图像特征之间的差异.该过程域迁移网络整体损失函数如公式(1)所示,$L_{reg}$为回归损失(即所估计的三维姿态与真实三维姿态之间的距离),$L_{domain}$为域迁移损失(分两阶段训练,其中,第1阶段固定特征提取器参数,目标是获得不错的姿态回归,并能够区分渲染图像与真实图像;第2阶段固定域间分类器的参数,目标是获取新特征,混淆域间分类器).

$$Loss = L_{reg} + L_{domain} \tag{1}$$

$$L_{reg} = \sum p_{pre} - p_{gt2} \tag{2}$$

$$L_{domain}^{stage1} = -\sum_{x \in real} Log_{p_x} - \sum_{x \in synthetic} Log_{(1-p_x)} \tag{3}$$

$$L_{domain}^{stage2} = -\sum_{x} (0.5 \times Log_{p_x} + 0.5 \times Log_{(1-p_x)}) \tag{4}$$

© 中国科学院软件研究所 http://www.jos.org.cn

(3) 人体三维骨架提取

对于人体三维骨架提取,我们采取 AlexNet[35]网络结构,利用生成的渲染数据及人体三维姿态坐标去训练新的模型.为了使现有网络结构适应于人体三维姿态估计任务,我们修改了这些网络的最后一层,使其能够直接输出三维坐标,并在推断的三维骨架和真实三维姿态中间添加一层欧几里德损失函数(见公式(5)),在训练过程中对全连接层进行微调以使得参数从一个良好的初始值去适应新的面向手术室的人体三维骨架提取任务.

$$E=\sum(P_i-Q_i^2) \tag{5}$$

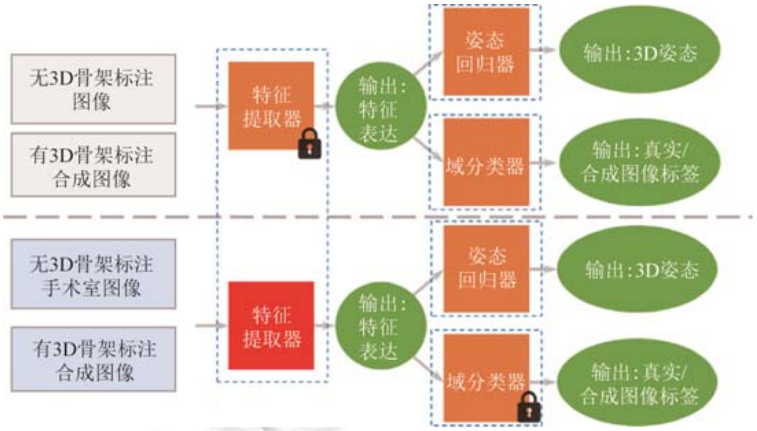



Fig.2 Occlusion-oriented skeleton extraction domain migration network

图 2 面向遮挡添加下骨架提取的域迁移网络

### 4.2 交互手势理解

为了准确理解无菌条件下的医生交互的手势,本文在一般改进的 DTW 算法的基础上,采用基于后验处理的优化方式,该方式通过参数调控、无效区域判定以及静止手势处理对 DTW 的输出结果进行修正,加快 DTW 的执行速度,将混淆手势作为无效手势处理,提升了手势识别率,并能够实现 DTW 算法无法处理的连续重复手势的识别,在实时性方面亦有较好的表现.

(1) 手势特征提取

Kinect 骨骼系统提供了 20 个关节点的三维坐标信息,如果将所有关节点都作为特征点,计算会过于复杂,关节点之间也会相互干扰.因此,本方法舍去了一些在手势序列中作用不明显的骨骼关节点,降低了计算的复杂度,提高了识别速度.定义了 7 个常用操作手势,分别为右手向右滑动、右手向左滑动、右手向上滑动、右手向下滑动、双手向外扩张、双手向内收拢、左手向左滑动.在这 7 种手势中,最重要的参考节点为右手关节、右肘关节、左手关节、左肘关节、双肩中心以及脊柱中央这 6 个节点.记第 $t$ 帧编号为 $i$ 的关节点的坐标为 $C_{i,t}=(X_{i,t},Y_{i,t},Z_{i,t})$,$X_{i,t},Y_{i,t},Z_{i,t}$ 分别表示第 $t$ 帧编号为 $i$ 的关节点在以 Kinect 为原点的三维坐标系下的 $x,y,z$ 的值.

由于在手势操作过程中变化的点仅为右手关节、右肘关节、左手关节、左肘关节,因此以这 4 个节点作为特征向量,双肩中心及脊柱中央节点作为参考节点,第 $t$ 帧的特征向量 $S$ 可表示为

$$S_t=(C_{rh,t},C_{re,t},C_{lh,t},C_{le,t}) \tag{6}$$

式(6)中,$C_{rh,t},C_{re,t},C_{lh,t},C_{le,t}$ 分别表示第 $t$ 帧右手、右肘、左手、左肘的坐标.Kinect 提供的关节点坐标对应以 Kinect 为原点的坐标系,在实际应用中,人体的高矮胖瘦以及相对 Kinect 的不同站立位置都会影响手势特征向量的具体数值.对节点坐标进行中心化和归一化可以有效地解决这一问题.以人体脊柱中央为坐标原点,脊柱中央到双肩中心的距离作为参考距离,进行中心化后,第 $t$ 帧编号为 $i$ 的关节点坐标变为

$$C'_{i,t}=(x_{i,t}-x_{sp,t},y_{i,t}-y_{sp,t},z_{i,t}-z_{sp,t}) \tag{7}$$

式(7)中,$x_{sp,t},y_{sp,t},z_{sp,t}$ 为脊柱中央节点的坐标值.脊柱中央到双肩中心的欧式距离为

© 中国科学院软件研究所 http://www.jos.org.cn

$$H_t=\sqrt{(x_{sc,t}-x_{sp,t})^2+(y_{sc,t}-y_{sp,t})^2+(z_{sc,t}-z_{sp,t})^2} \tag{8}$$

式(8)中,$x_{sc,t},y_{sc,t},z_{sc,t}$将节点坐标进行归一化,记归一化后第 $t$ 帧编号为 $i$ 的节点坐标为

$$G_{i,t}=\frac{C'_{i,t}}{H_t} \tag{9}$$

则第 $t$ 帧进行中心化和归一化的特征向量 $V$ 可表示为

$$V_t=(G_{rh,t},G_{re,t},G_{lh,t},G_{le,t}) \tag{10}$$

(2) 手势模板序列的训练

本文采用 DTW 进行手势模板序列的训练.DTW 算法的核心是将测试序列与模板序列进行匹配,因此,手势模板的选择将会很大程度上影响匹配的结果.本文用如下方法来确定手势模板序列,每个模板序列长度均为 20 帧.设样本序列 $k=(V_1,V_2,V_3,\ldots,V_{20})$,根据已经定义的 7 种手势动作,每种手势采集 $n$ 个样本 $K=(V_1,V_2,V_3,\ldots,V_m,\ldots,V_n)$.对于每个样本 $k_m$,依次与其余 $n$–1 个样本使用 DTW 进行匹配.记待测样本 $k_m$ 与样本 $k_1,k_2,k_3,\ldots$之间的 DTW 距离为 $d_1,\ldots,d_m$,则待测样本 $k_m$ 的累计规整距离为 $D_m=\sum_{i=1}^{n}d_m(i\neq m)$, 然后对每个手势类别下的样本进行计算,便可确定所有类别手势的模板序列.

$$D_T=\min(D_1,D_2,D_3,...,D_n) \tag{11}$$

将每一个样本均使用式(11)计算其累计规整距离 $D_1,D_2,D_3,\ldots,D_n$,累计规整距离越小,说明样本的代表性就越强,并以此作为确定所有类别手势的模板依据.

### 4.3 基于麦克风阵列的远场语音识别

针对手术室环境这一特殊的应用场景,因为医生难以通过手持麦克风直接进行语音交互,同时,头戴式麦克风目前也不是国内外手术室的基本配置,因此需要选择麦克风阵列作为拾音设备,采集不同方位的语音进行增强处理,在此基础上识别音频中的内容.本文采用这种端到端的建模方法以提高手术室这种复杂环境下语音识别的性能,从而实现在手术室环境下,医生能够释放双手进行语音交互.

(1) 语音前端处理

语音前端处理模块的顺序是回声消除、混响消除、波束形成、增益控制,然后在此基础上进行后端处理,接下来介绍采用这种顺序的原因:回声消除模块有参考信号源(比如远端喇叭播放的手术控制指令)可以参考,通过回声消除模块可以剔除远端信号的干扰,远端信号的干扰(比如播放手术控制指令)会影响到混响消除和波束形成算法的性能,因此,对于每一路麦克风,首先进行回声消除以消除其中一个干扰源的影响.在此基础上进行混响消除,混响消除放到波束形成之前的原因是混响与房间的特性相关,不同麦克风之间的关系可以反映出这种空间特性,因此,采用多通道混响消除方法;然后对多通道信号进行波束形成,生成单通道的信号;再对波束形成后生成的单通道信号进行后置滤波,消除残留噪声的干扰.声音在传输过程中可能会存在能量的衰减和溢出,通过增益控制算法对能量进行控制,生成最终经前端处理后输出的语音,用于后端语音识别或指令词识别的处理.语音前端处理流程如图 3 所示.

(2) 语音后端建模

面向手术室环境的语音识别系统由声学模型训练模块、语言模型训练模块和超大空间解码 3 个相互制约的部分组成;声学模型训练模块通过深度学习方法提升语音识别器的声学模型的泛化能力;语言模型训练模块通过融合 Grammer 和 $N$-gram 信息的方法在大规模数据集下训练鲁棒的语言模型.超大空间解码子系统针对战场环境的特点,通过高效约简的解码算法,快速、有效地从复杂搜索空间中确定最优路径,保证语音识别器的准确率和运行速度.语音识别系统能够支持在线对声学模型和语言模型更新,从而提高对特定环境的适应能力.多通道语音识别流程如图 3 所示,联合通用领域的声学模型和面向手术室环境的解码网络进行语音识别,通过并行训练方法训练基于深层神经网络的声学模型,通过迁移学习机制,实现对领域知识的更新,快速构建面向手术室环境的解码网络,面向手术室环境域的语音识别解码方案如图 4 所示.

© 中国科学院软件研究所 http://www.jos.org.cn

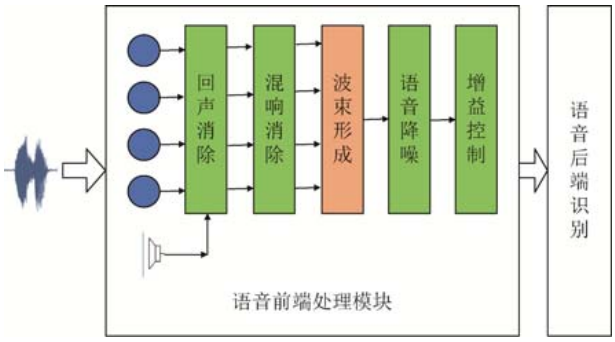



Fig.3 Far-field speech recognition front-end flow operating

图 3 远场语音识别前端处理流程

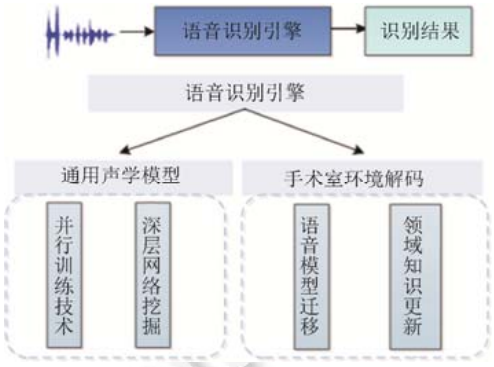



Fig.4 Speech recognition decoding for processing room environment domain

图 4 面向手术室环境域的语音识别解码

## 5 多通道信息融合方法

无菌手术环境中,在非接触式的自然交互情况下,由于语音识别的错误、姿态、手势受到遮挡,因此,交互系统难以统一单一模态信息,精确地判断医生的操作意图.为了提高交互系统中对医生意图识别的准确率,我们将多模态信息融合的不同策略引入神经网络模型.随着计算机技术和深度学习的快速发展,结构更深的神经网络模型在语音识别、人机对话、机器翻译、语义理解、目标识别、手势检测与跟踪、人体检测与跟踪等领域得到广泛应用.如在情感识别领域,采用相似度评估,目前采用深度长短时记忆神经网络模型(long short-term memory neural network,简称 LSTM)由计算机运行后得到的最好结果与专业人士识别相差 10%左右[64,65];在语音识别领域,目前针对方言口音的语音识别,深度递归神经网络(recurrent neural networks,简称 RNN)在字识别准确度上可以达到 95%[66],接近人类水平;在图像目标识别领域,超大规模深度卷积神经网络(convolution neural network,简称 CNN)已经超过普通人类辨识水平[67,68].深度神经网络模型技术在单一通道的数据处理上已经取得很好的成效,但是,如何构建面向多通道信息融合的大规模深度神经网络模型,在融合阶段无差别地处理多通道信息仍然是目前研究的热点问题.

为了更为精准地实现交互系统在手术室环境中对医生意图的识别,将多模态信息应用于深度神经网络,考虑到不同通道图像、语音、手势、生理信息的差异性,因此,在融合结构上,通过在特征层进行融合,具体的融合策略的抽象表示如图 5 所示.

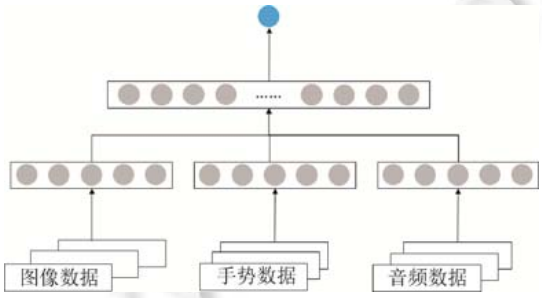



Fig.5 Multi-modal information fusion for operating room environment

图 5 面向手术室环境的多模态信息融合

图像数据主要是用于手术室复杂场景下人体的骨架提取,采用深度学习端对端特性从单张二维图像中全自动地提取三维人体骨架特征,该特征包含了 54 维参数的人体骨架特征;手势数据对应为 Kinect 获取的手势信

© 中国科学院软件研究所 http://www.jos.org.cn

息,由于手势操作主要是通过观察人手的右手关节、右肘关节、左手关节、左肘关节而实现,故我们将这 4 个节点作为特征向量,双肩中心及脊柱中央节点作为参考节点,构建 20 维特征向量作为手势特征;将采用麦克风阵列进行远场语音识别获取的数据作为音频数据,通过构建端到端的建模方法在每帧的音频数据中提取 64 维的音频特征.在多通道信息特征提取的过程中,由于不同通道信息数据采集的频率并不相同,因此需要对不同通道的数据进行不同的采样,并加以特征融合,构建融合特征向量,然后采用深度学习的方法对特征向量进行分类,以判断当前状态下医生的意图.

## 6 实 验

### 6.1 人体骨架结果及分析

#### 6.1.1 定量结果分析

(1) 人体三维骨架提取结果分析

如前文所述,训练数据集的好坏直接影响到卷积神经网络提取人体三维骨架的性能,该算法的核心贡献在于提出了一个大规模人体三维骨架数据集.为此,我们用不同的标准化卷积神经网络模型(Li14[28]、AlexNet 以及 VGG[69]),分别在经典 Human3.6M 数据集[70]、我们的数据集以及二者混合这 3 个数据集上对人体三维骨架网络进行训练,并在 Human3.6D+测试集评估各种方法及数据对应的人体三维骨架提取性能.如图 6 中左图所示结果,用本文数据集训练的模型要优于用 Human 3.6M 数据集训练的模型;Human3.6D+数据测试集合中图像的变化更为丰富,表明本文所合成的数据集能够更好地训练模型学习这些变化.

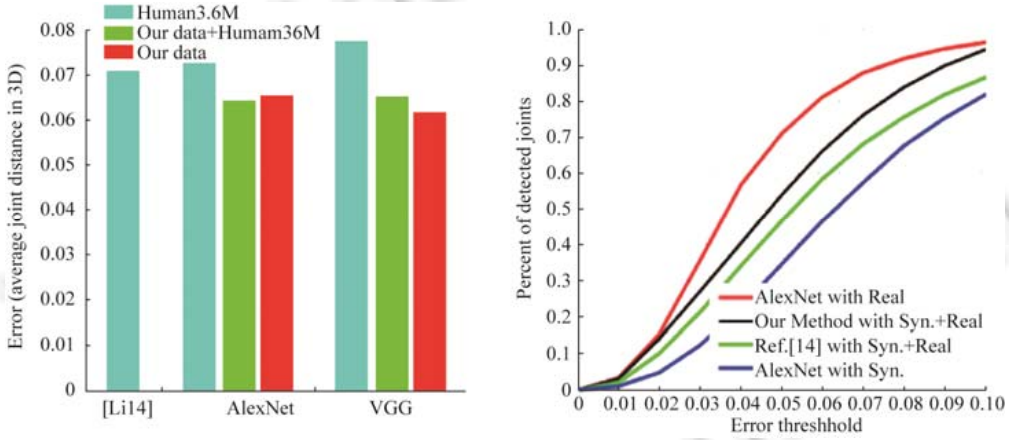



Fig.6 Quantitative results analysis

图 6 定量结果分析

图 6 的左图分析了在运用不同卷积神经网络模型的情况下,本文方法与 Human3.6M 数据集在 Human3D+测试集上生成的人体三维骨架测试结果;右图为域迁移网络结果分析,通过对不同方法使用混合数据(合成数据以及/或者真实数据)分析域迁移网络性能.可以看到,本文提出的域迁移网络不需要真实图像的三维人体姿态标签,因此,在训练过程中可通过添加大量真实图像抑制过拟合现象.如图 6 中右图所示,经过域迁移网络训练的合成图像及三维姿态,其模型在基准卷积神经网络模型上有着极大的提升,仅次于使用真实图像及真实三维姿态的结果,并且,网络结构明显优于经典域迁移网络[71].其原因在于,域迁移网络能够训练出更好的特征提取器,从真实图像和合成图像中提取出更为有意义的特征.

#### 6.1.2 定性结果分析

利用深度学习由二维图像生成三维人体骨架,核心在于构建大规模二维图像与相对应的三维人体姿态标注数据集.因此,我们构建了 Human3D+数据库,该数据库包含 1 574 幅丰富的人体运动动作二维图像及三维人

© 中国科学院软件研究所 http://www.jos.org.cn

体姿态坐标,能够较好地描述真实图像的分布.除此以外,我们根据现有人体三维骨架数据库中的姿态数据,合成了 10 556 个具有独特纹理及姿态的人体模型,经过背景与光照渲染,合成 5 099 405 幅训练图像用于网络训练.

对于单幅二维图像,可利用本文提出的算法获取三维人体骨架.图 7 展现了一组从单幅图像生成三维人体骨架的实验室结果,其中第 1 列和第 3 列为输入图像叠加了人体骨架的图像(圆球为人体主要关节点),第 2 列和第 4 列为利用现有算法(如 SCAPE 模型)匹配并重建的三维模型.

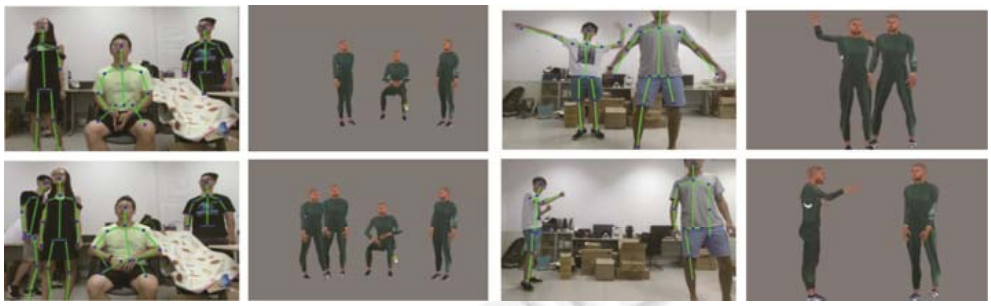


Fig.7 Extraction effect of multi-human skeleton with partly occlusion

图 7 遮挡条件下的多人骨架提取效果图

### 6.2 手势提取结果及分析

#### 6.2.1 正确性验证

为了验证本文基于后验处理的 DTW 优化方法仍具有较高的可行性,首先对孤立手势识别的正确率进行验证.本系统定义了 7 种操作手势,图 8 展示了实验所用手势的示意图.每幅图片右上方显示出对应的手势动作,“NoGesture”表示手势落在无效区域内.


Fig.8 Gesture schematics and invalid areas

图 8 手势示意图及其无效区域

使用传统 DTW 算法、文献[42]提出的改进 DTW 算法和本文基于后验处理的 DTW 优化方法分别进行实验,按照识别出来的独立手势名称进行统计,其混淆矩阵对比情况见表 1.

由混淆矩阵对比可以看出,传统 DTW 算法在进行连续重复手势处理时,无法区分正确手势与无关手势,从而导致系统执行了非常多的错误指令,其指令正确率几乎都低于 50%,对于双手动作的指令正确率更是低至 42%.文献[42]提出的改进 DTW 方法对连续重复手势的处理效果比传统 DTW 略有提升,基于后验处理的 DTW

© 中国科学院软件研究所 http://www.jos.org.cn

优化方法对于连续重复手势的处理效果则要好得多,指令正确率普遍高于 96%.使用 *G*-Mean 指标作为识别结果好坏的评价标准,分别用指令正确率 IA 和识别率 RR 代替式(24)中的召回率 REC,则可以得到“指令正确率”和“识别率”的 *G*-Mean 值,以此作为综合指令正确率和综合识别率.经计算后,传统 DTW 算法的综合指令正确率、综合识别率分别为 51.18%和 87.14%,文献[42]提出的改进 DTW 方法的综合指令正确率、综合识别率分别为 56.75%和 89.04%,基于后验处理优化的 DTW 算法综合指令正确率、综合识别率分别为 98.56%和 97.12%.实验结果表明,本文提出的后验处理优化方法能够有效识别用户的连续重复手势,在指令正确率和识别率上都优于传统 DTW 算法.

**Table 1** Comparison of confusion matrices for continuous gesture recognition based on two methods

**表 1** 两种方法的连续手势识别混淆矩阵对比

| 传统DTW | 右手连续右滑 | 右手连续左滑 | 右手连续上滑 | 右手连续下滑 | 双手连续扩张 | 双手连续收缩 | 左手连续左滑 |
|---|---|---|---|---|---|---|---|
| 右手右滑 | 45 | 44 | 0 | 0 | 4 | 3 | 0 |
| 右手左滑 | 44 | 44 | 0 | 0 | 6 | 5 | 0 |
| 右手上滑 | 2 | 0 | 46 | 45 | 3 | 2 | 0 |
| 右手下滑 | 0 | 3 | 44 | 46 | 2 | 3 | 0 |
| 双手扩张 | 1 | 0 | 3 | 0 | 38 | 40 | 1 |
| 双手收缩 | 1 | 3 | 0 | 2 | 38 | 39 | 0 |
| 左手左滑 | 0 | 0 | 0 | 0 | 0 | 0 | 48 |
| 无效手势 | 3 | 2 | 4 | 3 | 5 | 5 | 1 |
| 指令正确率 | 48% | 47% | 49% | 49% | 42% | 43% | 94% |
| 识别率 | 90% | 88% | 92% | 92% | 76% | 78% | 96% |

| 文献[42]改进DTW | 右手连续右滑 | 右手连续左滑 | 右手连续上滑 | 右手连续下滑 | 双手连续扩张 | 双手连续收缩 | 左手连续左滑 |
|---|---|---|---|---|---|---|---|
| 右手右滑 | 47 | 36 | 0 | 0 | 4 | 3 | 0 |
| 右手左滑 | 32 | 46 | 0 | 0 | 6 | 5 | 0 |
| 右手上滑 | 2 | 0 | 46 | 37 | 3 | 2 | 0 |
| 右手下滑 | 0 | 3 | 34 | 48 | 2 | 3 | 0 |
| 双手扩张 | 1 | 0 | 3 | 0 | 38 | 31 | 0 |
| 双手收缩 | 1 | 3 | 0 | 2 | 32 | 39 | 0 |
| 左手左滑 | 0 | 0 | 0 | 0 | 0 | 0 | 49 |
| 无效手势 | 8 | 6 | 7 | 6 | 10 | 7 | 1 |
| 指令正确率 | 57% | 52% | 55% | 55% | 45% | 47% | 100% |
| 识别率 | 94% | 92% | 92% | 96% | 76% | 78% | 98% |

| 后验处理优化方法 | 右手连续右滑 | 右手连续左滑 | 右手连续上滑 | 右手连续下滑 | 双手连续扩张 | 双手连续收缩 | 左手连续左滑 |
|---|---|---|---|---|---|---|---|
| 右手右滑 | 48 | 0 | 0 | 0 | 0 | 0 | 0 |
| 右手左滑 | 0 | 49 | 0 | 0 | 0 | 1 | 0 |
| 右手上滑 | 0 | 0 | 48 | 0 | 2 | 0 | 0 |
| 右手下滑 | 0 | 0 | 0 | 50 | 0 | 0 | 0 |
| 双手扩张 | 0 | 1 | 1 | 0 | 48 | 0 | 0 |
| 双手收缩 | 0 | 0 | 0 | 0 | 0 | 47 | 0 |
| 左手左滑 | 0 | 0 | 0 | 0 | 0 | 0 | 50 |
| 无效手势 | 49 | 47 | 50 | 49 | 49 | 5 | 49 |
| 指令正确率 | 100% | 98% | 98% | 100% | 96% | 98% | 100% |
| 识别率 | 96% | 98% | 96% | 100% | 96% | 94% | 100% |

6.2.2 实时性检测

5 名志愿者依次做一组由 7 个手势随机组合的动作,编号为序列 1、序列 2、序列 3、序列 4、序列 5.记录每个手势开始和首次识别成功时刻对应的帧编号,求其差值便可计算出识别每个手势所用的时间.表 2 记录了 5 个序列进行测试时每个手势从开始到成功识别经过的帧数及估计时间.

**Table 2** The timeliness of real-time gesture recognition based on posterior processing DTW optimization method

**表 2** 本文基于后验处理的 DTW 优化方法进行实时手势识别的时效性

| | 右手右滑 | 右手左滑 | 右手上滑 | 右手下滑 | 双手扩张 | 双手收缩 | 左手左滑 |
|---|---|---|---|---|---|---|---|
| 序列 1 | 6 | 7 | 8 | 5 | 6 | 11 | 5 |
| 序列 2 | 8 | 5 | 8 | 9 | 7 | 8 | 7 |
| 序列 3 | 7 | 5 | 7 | 6 | 7 | 9 | 5 |
| 序列 4 | 7 | 6 | 7 | 6 | 6 | 7 | 6 |
| 序列 5 | 6 | 6 | 6 | 8 | 5 | 9 | 7 |
| 平均值 | 6.8 | 5.8 | 7.2 | 6.8 | 6.2 | 8.8 | 6 |
| 估计时间(ms) | 224.4 | 191.4 | 237.6 | 224.4 | 204.6 | 290.4 | 198.0 |

实验结果表明,本文提出的基于后验处理的 DTW 优化方法可在大约 200ms~300ms 的时间延迟内给出识别结果并控制系统进行相关操作,能够满足识别实时性的要求.

### 6.3 语音信息处理结果及分析

(1) 实验数据

本文采用实测数据集进行实验结果评估,该数据集在手术室环境下实际录制,测试集中共包括 2 000 句样本,包括 100 个说话人,平均信噪比为 5dB,平均混响时间为 300ms;训练集采用仿真生成的远场数据进行训练,包括 2 000 小时的训练数据,信噪比覆盖 0dB、5dB、10dB、15dB,混响时间涉及 100ms、200ms、300ms、400ms 和 500ms.麦克风阵列设备采用 6+1 的环形阵列.测试样本主要包括医院手术相关命令词汇,如“开始手术”“打开设备”“到第 8 页”“监控心电图”“准备麻醉”等.

(2) 实验设置

© 中国科学院软件研究所 http://www.jos.org.cn

本文在语音识别工具 Kaldi 的基础上进行开发和实验,实验共采用两种特征:mel 频率倒谱系数(MFCC)和 mel 标度滤波器组特征(FBANK).提取特征的窗长为 25ms,帧移为 10ms.MFCC 特征为 13 维,加上其一阶和二阶差分统计量,共 39 维.FBANK 特征为 40 维,加上其一阶和二阶差分统计量,共 120 维.特征的均值方差归一化以说话人为单位进行.所有 GMM-HMM 的输入为 MFCC,所有神经网络模型的输入为 FBANK.就本文所涉及到的神经网络模型而言,其损失函数为交叉熵,优化准则为随机梯度下降(SGD).DNN 模型采用反向传播(BP)算法进行训练.BLSTM 模型采用随时间反向传播(BPTT)算法进行训练.LSTM 模型采用截断的随时间反向传播(truncatedBPTT)算法进行训练.本文实验所用语言模型为三元文法语言模型,词表大小为100G;解码的搜索空间基于加权有限状态转换器(WFST)进行构建,搜索策略为束搜索(beam-search)算法.

(3) 基线方法

基线方法中前端采用加权延时求和方法进行增强处理,后端分别采用DNN和LSTM-RNN进行声学模型训练;所有 DNN 模型均含有 7 个隐层,每个隐层含有 2 048 个节点.LSTM-RNN 模型含有 5 个隐层,每个隐层包含 640 个单元.DNN 模型的初始学习速率为 0.008,LSTM-RNN 的初始学习速率为 0.000 01,冲量值均设为 0.9.

(4) 实验结果对比

本文前端波束形成采用广义旁瓣滤波方法,去混响采用加权预测误差方法,采用深层神经网络进行单通道语音增强处理.本文采用的方法将 DNN 和 LSTM-RNN 两种声学模型输出的后验概率进行融合,通过联合建模的方式提高语音识别的性能.实验结果见表 3.

**Table 3** Comparison of speech recognition experiments

**表 3** 语音识别实验结果对比

| 方法 | 性能(%) |
|---|---|
| 延时求和+DNN | 82.36 |
| 延时求和+LSTM-RNN | 84.25 |
| 延时求和+模型融合 | 85.56 |
| 本文前端+DNN | 87.64 |
| 本文前端+LSTM-RNN | 88.56 |
| 本文前端+模型融合 | 89.37 |

针对“本文前端+模型融合”的模型,在不同距离下进行了语音识别实验,实验结果见表 4.

**Table 4** Comparison of speech recognition under different distances

**表 4** 不同距离条件下语音识别实验结果对比

| 距离(m) | 性能(%) |
|---|---|
| 1 | 91.37 |
| 2 | 89.44 |
| 3 | 87.52 |

(5) 实验结果分析

通过对比表 3、表 4 中的实验结果可知,语音前端处理对于提升语音识别的性能起着非常关键的作用,本文采用的广义旁瓣滤波方法通过自适应波束形成可以有效地增强目标方向的声音,同时,通过加权预测误差消除了远场语音的干扰,在此基础上,通过深层神经网络模型有效地消除了非平稳噪声的干扰,因此,相比于延时求和这种固定波束形成方法,有效地提升了语音识别的性能.同时,本文采用的模型融合策略,可以有效提升声学模型的建模精度,融合后的模型结合了 DNN 和 LSTM-RNN 两种模型的优势,从而提升了语音识别在真实环境下的鲁棒性.

### 6.4 无菌条件非接触式多通道自然交互手术环境信息融合结果及分析

根据各单一通道技术与融合的要求,本文设计并接近真实地构建了整个无菌条件非接触式多通道自然交互手术环境,设计时,要充分考虑各单一模态和融合计算的需求.系统的各个组成部件选用标准的硬件和软件,采用模块化设计,使系统可以通过增加模块的方式进行扩容.无菌条件非接触式多通道自然交互手术环境整体

© 中国科学院软件研究所 http://www.jos.org.cn

布局的要求如下:(1) 节约手术室空间,使手术室更为简洁,便于远场语音信息采集及姿态和手势获取;(2) 采用四分屏 50 吋显示器可实现阅片,显示监护仪、内窥镜等设备的图像,方便获取手术信息;(3) 全景摄像方便手术室内场景实时监控;(4) 嵌入式一体化工作站节约空间,双屏设计,便于操作;(5) 双 26 吋内窥显示器,可用于内窥镜手术场景.图 9 给出了无菌条件非接触式多通道自然交互手术环境设计与真实场景图.


Fig.9 Aseptic sonditional contactless multichannel natural interaction durgery environment design (left) and real scene map (right)

图 9 无菌条件非接触式多通道自然交互手术环境设计(左)与真实场景图(右)

在医院手术环境的交互过程中,交互系统对医生的意图识别的准确度和速度十分重要.根据无菌条件非接触式多通道自然交互手术的设计及搭建的真实场,本文设置了 10 种自然的医生手势动作,根据在不同操作视框的定义,可以实现呈线性倍数数量的指令,完全满足系统的交互模式.本实验通过对比姿态、手势、语音等单一通道信息和多通道信息融合条件下系统对医生意图识别的准确度和速度,并分析在不同单一通道和多通道融合对医生意图识别的影响,发现在交互过程中,虽然基于单一通道的手势、语音信息能够使得系统在对医生意图识别时取得较好的准确度和速度,但是相对而言,基于多通道信息融合的效果会更好.实验结果见表 5 和表 6.

**Table 5** Accuracy of doctor intention recognition based on single channel information and multi-channel information fusion (%)

**表 5** 单一通道信息和多通道信息融合对医生意图识别准确率(%)

| | 手势 | 语音 | 手势+语音 |
|---|---|---|---|
| 确定 | 94.52 | 92.46 | 95.32 |
| 上一张 | 93.82 | 93.51 | 96.74 |
| 下一张 | 94.67 | 93.75 | 96.38 |
| 上移 | 82.58 | 92.86 | 95.18 |
| 下移 | 83.65 | 92.73 | 96.59 |
| 左移 | 84.52 | 93.65 | 95.48 |
| 右移 | 82.65 | 91.26 | 94.81 |
| 放大 | 80.52 | 93.25 | 96.27 |
| 缩小 | 79.62 | 91.25 | 95.16 |
| 返回 | 80.65 | 93.45 | 96.46 |
| 执行特定操作平均准确度 | 85.72 | 92.81 | 95.83 |

从实验结果可以看出,单一通道条件下,由于手势信息较为复杂,但手术室环境噪声较小,所以系统在基于手势交互的基础上对医生意图的识别率比语音较低,但在某些较为简单的手势动作上,如“确定”“上一张”“下一张”的准确度并不比语音信息差.而在识别的时间方面,较为复杂的手势动作同样不占优势,但是对简单手势动作而言,它们的识别速度仍然比语音来得更快.在此基础上,将不同通道的信息进行融合后,系统无论是在时间性能上,还是在意图理解的准确度上都会有比较明显的提升.并且,相较于通过授意护士或者手术助理到计算机操作室操作的方式(以乳腺癌肿瘤手术为例,护士或者手术助理到计算机操作室定位到病灶图像平均约 1 分钟),本文的定位方法平均不超过 2s,可以看到,采用多通道信息融合方式来识别医生的意图,可以更快地定位到病灶图像.

© 中国科学院软件研究所 http://www.jos.org.cn

由于在手术室的操作环境中,多通道信息的融合处理的结果主要是为医生在手术时提供便捷的交互环境,因此在交互过程中,医生对多通道信息融合结果的满意程度也很重要.因此,我们邀请了 4 位医生以及 32 位助理人员对多通道信息融合交互系统的结果进行体验和评测,每人至少进行 3 轮以上的有效操作,最后在其他评测结束后,要求每个体验医生对结果进行满意度投票,总共 5 个选项,分别是很满意、满意、一般、不太满意和很不满意,其统计分布如图 10 所示.

由图 10 所示评测结果可知,66.67%的测试人员对多通道信息融合的结果体验感觉满意或者很满意,而只有 16.67%的医生对体验不太满意或者很不满意.从用户的主观评测角度来看,医生对多通道信息融合的交互体验比较不错,能够获得大多数体验医生的认可.

**Table 6** Speed of doctor intention recognition based on single channel information and multi-channel information fusion

**表 6** 单一通道信息和多通道信息融合对医生意图识别的速度

| 速度(ms) | 手势 | 语音 | 手势+语音 |
|---|---|---|---|
| 确定 | 865 | 1 085 | 670 |
| 上一张 | 953 | 1 152 | 852 |
| 下一张 | 1 023 | 956 | 753 |
| 上移 | 856 | 1 026 | 695 |
| 下移 | 962 | 1 075 | 716 |
| 左移 | 982 | 982 | 829 |
| 右移 | 849 | 1 012 | 658 |
| 放大 | 1 356 | 968 | 743 |
| 缩小 | 1 428 | 1 049 | 675 |
| 返回 | 1 185 | 912 | 852 |
| 执行特定操作平均时间 | 1 045.9 | 1 021.7 | 744.3 |

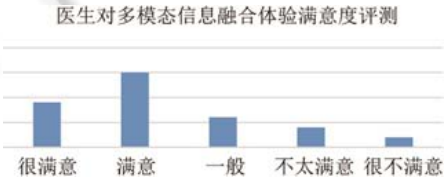



Fig.10 Evaluation of doctors for multimodal information fusion experience

图 10 医生对多模态信息融合体验满意度评测

## 7 结论和展望

实验结果表明,在接近实际的实验环境中,通过融合遮挡环境下的深度图像人体骨架提取、手势跟踪与理解、手术室环境远场语音识别,多模态信息处理与融合技术,无菌条件下的非接触式多通道自然交互手术环境相对于传统的通过护士或者手术助理到计算机操作室操作病灶图像的方式,能够明显地节省时间,使得主治医生在手术中可通过语音命令、手势及上述交互相结合的方式快速定位到需要观察的病灶成像.本文建立的无菌条件的非接触式多通道自然交互手术环境在保证精度的情况下,为建立下一代未来高效的手术室提供了技术与方法验证,可极大地方便医生的手术过程,缩短平均手术时间.但无菌条件下的非接触式多通道自然交互手术环境距离把人机交互技术鲁棒地应用到临床还有一定距离,未来进一步的工作主要包括:(1) 进一步优化语音识别技术,更加准确地融合手势,更加准确地识别手术医师的意图;(2) 进一步引入三维手术影像导航技术,与多模态交互手段相融合,做到面向交互的更逼真的临床展示.

**References**:

[1] Wang HQ, Wang P, Wang F, *et al*. Design and practice of intelligent digital operating room system. Journal of Medical Intelligence, 2018,39(6):30–33 (in Chinese with English abstract).

© 中国科学院软件研究所 http://www.jos.org.cn

[2] Zhu C, Wu LY. Selection and implementation of digital operating room. Modern Hospital Management, 2014,12(6):77–80 (in Chinese with English abstract).

[3] Yang K, Cai YX, Fan PS, *et al*. Intelligent digital operating room design and implementation. Health Frontier, 2017,26(2) (in Chinese with English abstract).

[4] Hu YF. Intelligent operating room with gesture command. 2015 (in Chinese). http://hnrb.voc.com.cn/hnrb_epaper/html/2015-11/06/content_1029876.htm?div=-1

[5] Chorowski J, *et al*. Attention-based models for speech recognition. Computer Science, 2015.

[6] Wang L, *et al*. Recent developments in human motion analysis. Pattern Recognition, 2003,36:585–601.

[7] Betke CM. Fagiani, JG. Evaluation of tracking methods for human-computer interaction. In: Proc. of the IEEE Workshop on Applications in Computer Vision. 2002. 121–126.

[8] Hu W, *et al*. A survey on visual surveillance of object motion and behaviors. IEEE Trans. on Systems, Man, and Cybernetics, 2004,(34):334–352.

[9] Oudeyer PY. The production and recognition of emotions in speech: Features and algorithms. Int'l Journal of Human-Computer Studies, 2003,(59):157–183.

[10] Ge LH, *et al*. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proc. of the CVPR. 2017.

[11] Hasan HS, Kareem SA. Human computer interaction for vision based hand gesture recognition: A survey. Artificial Intelligence Review, 2015,(43):1–54.

[12] Ruffieux S, *et al*. A survey of datasets for human gesture recognition. In: Proc. of the Int'l Conf. on Human-computer Interaction. 2014. 337–348.

[13] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. In: Proc. of the Int'l Conf. on Computer Vision. 2017.

[14] Hatice G, Massimo P. Affect recognition from face and body: Early fusion vs. late fusion. In: Proc. of the IEEE Int'l Conf. on Systems, Man and Cybernetics. IEEE, 2006,(4):3437–3443.

[15] Yang MH, *et al*. A nature multimodal human-computer-interaction dialog system. In: Proc. of the CHCI in Harmonious Human Computer Environment, 2013 (CHCI 2013). 2013 (in Chinese with English abstract).

[16] Yang MH, *et al*. User behavior fusion in dialog management with multi-modal history cues. Multimedia Tools and Applications, 2015,(74):10025-10051.

[17] Ngiam JQ, *et al*. Multimodal deep learning. In: Proc. of the Int'l Conf. on Machine Learning. 2011. 689–696.

[18] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proc. of the 25th Int'l Conf. on Machine Learning, 2008. 160–170.

[19] Seltzer ML, Droppo J. Multi-task learning in deep neural networks for improved phoneme recognition. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. 2013.

[20] Huang FF, Cao JT, Ji XF. Two-human interaction recognition algorithm based on multi-channels information fusion. Computer Technology and Development, 2016,26(3):58–62 (in Chinese with English abstract).

[21] Mori G, Malik J. Recovering 3d human body configurations using shape contexts. TPAMI, 2006,28(7):1052–1062.

[22] Ferrari V, Marin-Jimenez M, Zisserman A. Progressive search space reduction for human pose estimation. In: Proc. of the CVPR. 2008. 1–8.

[23] Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. In: Proc. of the CVPR. 2014.

[24] Ye M, Wang XW, Yang RG, Ren L, Pollefeys M. Accurate 3D pose estimation from a single depth image. In: Proc. of the ICCV. 2011.

[25] Choi CH, Christensen HI. 3D pose estimation of daily objects using an RGB-D camera. In: Proc. of the IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. 2012.

[26] Fan X, Zheng K, Lin Y, Wang S. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proc. of the CVPR. 2015.

© 中国科学院软件研究所 http://www.jos.org.cn

[27] Gkioxari G, Hariharan B, Girshick R, Malik J. RCNNS for pose estimation and action detection. arXiv Preprint arXiv: 1406.5212, 2014.

[28] Li S, Chan AB. 3D human pose estimation from monocular images with deep convolutional neural network. In: Proc. of the ACCV. Springer-Verlag, 2014. 332–347.

[29] Belagiannis V, Wang X, Shitrit H, Hashimoto K, Stauder R, Aoki Y, Kranzfelder M, Schneider A, Fua P, Ilic S, Feuner H, Navab N. Parsing human skeletons in an operating room. In: Proc. of the Machine Vision and Applications. 2016.

[30] KadkhodamohammadiA, Gangi A, de Mathelin M, Padoy N. Articulated clinician detection using 3D pictorial structures on RGB-D data. Medical Image Analysis, 2017,35:215–224.

[31] Felzenszwalb PF, Huttenlocher DP. Pictorial structures for object recognition. Int'l Journal of Computer Vision, 2005,61(1):55–79.

[32] Kadkhodamohammadi A, Gangi A, de Mathelin M, Padoy N. A multi-view RGB-D approach for human pose estimation in operating rooms. In: Proc. of the WACV. 2017. 363–372.

[33] Ruan X, Tian C. Dynamic gesture recognition based on improved DTW algorithm. In: Proc. of the 2015 IEEE Int'l Conf. on Mechatronics and Automation (ICMA). Beijing, 2015. 2134–2138.

[34] He C, Hu ZF, Wang Y. Novel dynamic gesture recognition method based on improved DTW algorithm. Digital Communication, 2013.

[35] Pan H, Li J. Online human action recognition based on improved dynamic time warping. In: Proc. of the 2016 IEEE Int'l Conf. on Big Data Analysis (ICBDA). Hangzhou, 2016. 1–5.

[36] Hiyadi H, Ababsa F, Montagne C, Bouyakhf EH, Regragui F. Adaptive dynamic time warping for recognition of natural gestures. In: Proc. of the 6th Int'l Conf. on Image Processing Theory, Tools and Applications (IPTA). Oulu, 2016. 1–6.

[37] Zhang Z, Liu Y, Li A, *et al*. A novel method for user-defined human posture recognition using Kinect. In: Proc. of the Int'l Congress on Image and Signal Processing. IEEE, 2014. 736–740.

[38] Chen Y, Luo B, Chen YL, Liang G, Wu X. A real-time dynamic hand gesture recognition system using Kinect sensor. In: Proc. of the 2015 IEEE Int'l Conf. on Robotics and Biomimetics (ROBIO). Zhuhai, 2015. 2026–2030.

[39] Zhang Y, Wu S, Luo Y. Applications and recognition of gesture trajectory using HMM. Bandaoti Guangdian/Semiconductor Optoelectronics, 2015,36(4):650–656.

[40] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. of the IEEE, 1989,77(2): 257–286.

[41] Song Y, Gu Y, Wang P, *et al*. A Kinect based gesture recognition algorithm using GMM and HMM. In: Proc. of the Int'l Conf. on Biomedical Engineering and Informatics. IEEE, 2014. 750–754.

[42] Wu Q, Zhan Y, Shao Y, *et al*. Human motion matching and evaluation based on STDTW and *K*-means. Application of Electronic Technique, 2016.

[43] Li Y, Yang Y, Chen Y, Zhu M. A pre-training strategy for convolutional neural network applied to Chinese digital gesture recognition. In: Proc. of the 8th IEEE Int'l Conf. on Communication Software and Networks (ICCSN). Beijing, 2016. 620–624.

[44] Chavan P, Ghorpade T, Padiya P. Indian sign language to forecast text using leap motion sensor and RF classifier. In: Proc. of the 2016 Symp. on Colossal Data Analysis and Networking (CDAN). Indore, 2016. 1–5.

[45] Anguera X, Wooters C, Hernando J. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. on Audio Speech & Language Processing, 2007,15(7):2011–2022.

[46] Higuchi T, Ito N, Yoshioka T, *et al*. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE, 2016. 5210–5214.

[47] Warsitz E, Haeb-Umbach R. Blind acoustic beamforming based on generalized eigenvalue decomposition. IEEE Trans. on Audio Speech & Language Processing, 2007,15(5):1529–1539.

[48] Roux JL, Vincent E. Consistent wiener filtering for audio source separation. IEEE Signal Processing Letters, 2013,20(3):217–220.

[49] Nugraha AA, Liutkus A, Vincent E. Multichannel audio source separation with deep neural networks. IEEE/ACM Trans. on Audio Speech & Language Processing, 2015,24(9):1652–1664.

[50] Heymann J, Drude L, Haebumbach R. Neural network based spectral mask estimation for acoustic beamforming. IEEE Trans. on Industrial Electronics, 2016,46(3):544–553.

© 中国科学院软件研究所 http://www.jos.org.cn

[51] Lebart K, Boucher JM, Denbigh PN. A new method based on spectral subtraction for speech dereverberation. Acta Acustica United with Acustica, 2001,87(3):359–366.

[52] Yoshioka T, Nakatani T, Miyoshi M, *et al*. Blind separation and dereverberation of speech mixtures by joint optimization. IEEE Trans. on Audio Speech & Language Processing, 2010,19(1):69–84.

[53] Han K, Wang Y, Wang DL, *et al*. Learning spectral mapping for speech dereverberation and denoising. IEEE/ACM Trans. on Audio Speech & Language Processing, 2015,23(6):982–992.

[54] Gao J, Du J, Kong C, *et al*. An experimental study on joint modeling of mixed-bandwidth data via deep neural networks for robust speech recognition. In: Proc. of the IJCNN. IEEE, 2016. 588–594.

[55] Narayanan A, Wang DL. Joint noise adaptive training for robust automatic speech recognition. In: Proc. of the ICASSP. 2014. 2504–2508.

[56] Gao T, Du J, Dai LR, *et al*. Joint training of front-end and back-end deep neural networks for robust speech recognition. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE, 2015. 4375–4379.

[57] Wang Q, Du J, Bao X, *et al*. A universal VAD based on jointly trained deep neural networks. In: Proc. of the Interspeech. 2015.

[58] Liu B, Tao J, Zhang D, Zheng Y. A novel pitch extraction based on jointly trained deep BLSTM recurrent neural networks with bottleneck features. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. 2017.

[59] Li XW, *et al*. A priori knowledge accumulation and its application to linear BRDF model inversion. Journal of Geophysical Research-Atmospheres, 2001,106:11925–11935.

[60] Lake BM, *et al*. Human-level concept learning through probabilistic program induction. Science, 2015,350.

[61] Lehrmann AM, Gehler PV, Nowozin S. A nonparametric bayesian network prior of human pose. In: Proc. of the ICCV 2013. Washington: IEEE Computer Society, 2013. 1281–1288.

[62] Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J. SCAPE: Shape completion and animation of people. ACM ToG, 2005,24(3):408–416.

[63] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in Neural Information Processing Systems 25. Curran Associates, Inc., 2012. 1097–1105.

[64] Chao LL, *et al*. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proc. of the ACM Multimedia. 2015. 65–72.

[65] He L, *et al*. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proc. of the ACM Int'l Workshop on Audio/Visual Emotion Challenge. 2015. 73–80.

[66] Yu D, *et al*. Large-margin minimum classification error training for large-scale speech recognition tasks. In: Proc. of the IEEE Int'l Conf. on Acoustics. 2016.

[67] He KM, *et al*. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proc. of the IEEE Computer Vision and Pattern Recognition. 2015.

[68] Yang F, *et al*. Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2016. 2129–2137.

[69] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recoginition. In: Proc. of the ICLR. 2015.

[70] Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI, 2014,36(7):1325–1339.

[71] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. In: Proc. of the 32nd Int'l Conf. on Machine Learning (ICML 2015). 2015.

**附中文参考文献**:

[1] 王红迁,汪鹏,王飞,等.智能数字化手术室系统设计与实践.医学信息学杂志,2018,39(6):30–33.

[2] 朱晨,吴玲燕.数字化手术室的选型与实施.现代医院管理,2014,12(6):77–80.

[3] 杨琨,蔡亚欣,樊沛澍,等.智能数字化手术室整体设计与实施.健康前沿,2017,26(2).

[4] 胡宇芬.智能手术室,手势来指挥.2015.http://hnrb.voc.com.cn/hnrb_epaper/html/2015-11/06/content_1029876.htm?div=-1

[15] 杨明浩,等.面向自然交互的多通道人机对话系统.见:第 9 届全国和谐人机环境联合学术会议(CHCI 2013).2013.

© 中国科学院软件研究所 http://www.jos.org.cn

[20] 黄菲菲,曹江涛,姬晓飞.基于多通道信息融合的双人交互动作识别算法.计算机技术与发展,2016,26(3):58–62.

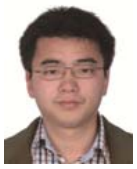
**陶建华**(1971－),男,江苏淮安人,博士,教授,博士生导师,CCF 会士,主要研究领域为语音合成,语音识别,情感计算,人机对话.

**杨明浩**(1977－),男,博士,副研究员,CCF 专业会员,主要研究领域为人机融合感知与决策,身自化认知计算理论与方法,多通道交互信息处理.

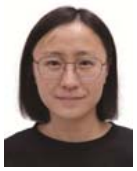
**王志良**(1956－),男,博士,教授,博士生导师,主要研究领域为人工心理理论与方法,机器人,物联网.

**班晓娟**(1970－),女,博士,正高级,博士生导师,CCF 高级会员,主要研究领域为人工智能,自然人机交互,三维可视化技术.

**解仑**(1968－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为情感计算与智能交互,智能机器人技术,工业控制系统信息安全.



**汪云海**(1984－),男,博士,教授,博士生导师,CCF 专业会员,主要研究领域为计算机图形学,数据可视化.



**曾琼**(1987－),女,博士后,CCF 专业会员,主要研究领域为计算机图形学,数据可视化.

**王飞**(1982－),男,高级工程师,主要研究领域为医疗信息化,医学大数据,人工智能.

**王红迁**(1987－),男,工程师,主要研究领域为大数据架构与研发,人工智能,医疗信息化.

**刘斌**(1984－),男,博士,副研究员,CCF 专业会员,主要研究领域为语音处理,虚拟听觉.

**韩志帅**(1993－),男,硕士,主要研究领域为人机交互,计算机视觉.

**潘航**(1991－),男,硕士,主要研究领域为情感计算,模式识别.

**陈文拯**(1992－),男,硕士,主要研究领域为计算机视觉.

© 中国科学院软件研究所 http://www.jos.org.cn