

基于视觉手势识别的人机交互系统

宋一凡 张 鹏 刘立波
(宁夏大学信息工程学院 银川 750000)

摘 要 人机交互系统是人与机器之间交流与信息传递的桥梁,随着计算机技术的迅速发展,使用鼠标、键盘等传统的人机交互技术已经不满足时代发展的需求,人们需要一种更快捷、更自然、更舒适的人机交互技术。基于手势的人机交互是人机交互系统的重要技术之一,传统的手势识别方法存在识别准确率不高、识别过程复杂等问题。针对上述缺陷,文中提出了一种基于深度学习的手势识别算法,该算法通过姿态估计对手势关节特征进行快速检测,利用卷积神经网络对关节特征图进行分类,克服了复杂背景中手势图像分割困难等问题,提高了识别结果的准确率。实验结果表明,该方法对各种手势不同尺度的表现具有很好的识别准确率,识别结果的准确率达到 98%。最后文中基于该算法设计了一个人机交互系统,并展示了手势识别在该人机交互系统中的应用。

关键词 姿态分析,深度学习,姿态预测,机器人,手势识别

中图分类号 TP242 **文献标识码** A

Human-machine Interaction System with Vision-based Gesture Recognition
SONG Yi-fan ZHANG Peng LIU Li-bo
(School of Information Engineering, Ningxia University, Yinchuan 750000, China)

Abstract Human-machine interaction system is the connection between human and computer. As the advances in computer technology, uses of mouse or keyboard are insufficient, now people need some nature and comfortable methods to manipulate the computers and so on. Gesture recognition-based method is one of the important human-machine interaction system, but there are some issues in the traditional method, such as low prediction accuracy, complex process procedure. To solve these problems, this paper proposed a deep learning-based gesture recognition algorithm. This method abstracts gesture feature heat map by pose estimation, then classified them by using convolutional neural network, overcomes the difficulty of image segmentation in complex background, and improves the accuracy in recognition. The result shows a 98% accuracy in recognition. Finally, a gesture guessing game robot was designed with this method, and the application of gesture recognition in human-machine interaction was presented.

Keywords Pose estimation, Deep learning, Pose prediction, Robot, Gesture recognition

1 引言

随着计算机技术与机器人技术的迅速发展,人机交互技术成为了计算机及机器人领域中极其重要的组成部分。人机交互是人与机器之间通过某种方式相互传递和交流信息,从而协同完成某种特定任务。经过多年的发展,人机交互技术取得了显著的进步,已成为计算机科学领域中不可或缺的一部分^[1]。

传统的人机交互一般是基于显示器、鼠标和键盘的交互方式,人通过鼠标、键盘或是其他输入设备对机器发送指令信息,机器再严格按指令行动,通过显示器向人传递反馈信息。然而,随着机器人技术的发展和智能机器人的出现,这类传统的基于键盘和鼠标的人机交互方式不再适用于现代人机交互系统,考虑到使用者可能并不熟悉电脑的操作,或者不方便使用鼠标或键盘对机器人下达指令,对此,需要一种更简单、更

便捷的交互方式,类似于人与人之间的自然交互方式^[2]。在人与人的交流当中,语言被公认是最自然、有效的交流方式,基于语音的人机交互获得了广泛的关注。在很多触控式操作不适用的场合,基于语音的人机交互系统发挥了不可替代的作用。比如驾驶员使用 GPS 导航系统导航时,语音交互是最方便的方式。对于一般的用户,语音交互也带来了良好的用户体验,基于语音的人机交互系统的应用已经十分普遍,如苹果的 Siri、微软的 Cortana 以及谷歌的 Google Assistant 都借助深度学习,在自然语言处理中表现优异,目前已被广泛的使用,在市场中的表现也十分成熟。然而人的交流不仅需要声音,还需要肢体、表情等更广泛的沟通方式,基于手势的人机交互成为了人机交互的重要部分。

除了语言外,手势交流也是人与人之间互动的重要方式。手势交流是一种肢体语言,可以自然直观的表达人类的情绪和意图。与语言不同,手势具有通用性,当一个人来到了一个

本文受自然科学基金(61862050),宁夏智慧农业关键技术集成应用与示范项目(2017BY067),宁夏回族自治区重点研发项目(2018BBF02006),社会科学基金(16BTY111)资助。

宋一凡(1989—),男,硕士生,主要研究方向为深度学习,E-mail:yifanwz@126.com;张 鹏(1975—),男,博士,副教授,CCF 会员,主要研究方向为智能信息处理,E-mail:pengzhang123@foxmail.com(通信作者)。

的置信图,用来预测所有关节的输出,每一层代表一个关节的输出,外加一个背景输出。

第 2 阶段到第 5 阶段的结构都相同,原始图像通过一个结构为由 3 个 9 × 9 卷积层构成和 1 个 5 × 5 卷积层构成的网络。得到的特征图连接上一阶段输出的置信图;再通过一个结构为由 3 个 11 × 11 卷积层和 2 个 1 × 1 卷积层构成的网络。最后输出 P + 1 层置信图,预测各关节输出。CPM 的每个阶段都输出所有关节的预测位置的热图,后一阶段对位置进行精炼,第 5 阶段得到最终关节特征热图,如图 3 所示。

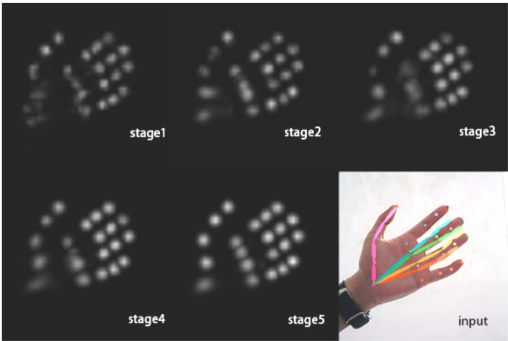


图 3 不同阶段的关节特征热图

为了避免梯度消失,图 2 中每一个阶段的输出都会使用一个损失函数将预测关节位置与理想关节位置的误差最小化。每一个关节位置的理想位置置信图记作:

$$b_p^*(Y_p)=Z$$

(1)

定义最小化误差的损失函数为:

$$f_i=\sum_{p=1}^{P+1}\sum_{z\in Z}\|b_i^p(z)-b_p^*(z)\|_2^2$$

(2)

其中, p 遍历所有关节, z 表示对应关节的位置,最后将所有阶段的损失函数累加,得到:

$$F=\sum_{i=1}^T f_i$$

(3)

所有阶段的网络参数均使用标准随机梯度下降法,在第 2 阶段及之后的所有阶段,通过共享相对应卷积层的网络权重参数,达到共享图像特征图的效果。

ECPM 的最后一个阶段是一个基本卷积神经网络,由 4 个卷积层和 4 个最大池化层组成,对第 5 阶段输出的关节特征热图进行特征提取,得到的结果再通过 3 个全连接层。最后的全连接层输出维度为 3,对应 3 种不同的手势,将每种手势图像映射到相应的类别。

3.2 预测模块

人的单次出拳是完全随机的,几乎无法准确预测,但是在一系列出拳中,一个人的出拳顺序是有某种倾向的。这种倾向性构成了一个具有规律的序列。

长短期记忆网络(Long-short Term Memory Network, LSTM)^[12]是一种改进的循环神经网络(Recurrent Neural Network, RNN),在 RNN 的基础上增加了一路输入、一路输出和遗忘门。增加的这一路输入和输出,被称为细胞状态,决定了保留和遗忘的信息,从而解决了 RNN 的长期依赖问题。LSTM 在多种序列处理任务中取得了最好的效果,包括语音识别和手写数字识别^[13]。LSTM 近年来发展尤其迅猛,在处理时间序列相关任务时的应用较为广泛。LSTM 网络结构及单个细胞内部结构如图 4 所示。

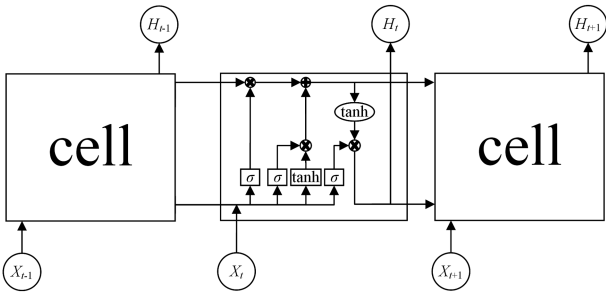


图 4 LSTM 网络结构节细胞内部结构

本文将人的动作序列表示为一个时间序列 $X\{x_1, x_2, \dots, x_t, \dots\}$ 。其中, x_t 代表动作执行过程中第 t 次的动作。用 0, 1, 2 代表石头、剪刀、布,构建 1 个 LSTM 网络,其包含 4 个隐藏层和 1 个输出层,时间步长 timestep 设置为 1。最终输出预测值 0, 1 或 2,表示预测下一次出拳的类别。激活函数使用 tanh,以动作序列作为数据集,损失函数使用均方误差(Mean-Square Error, MSE),均方误差的表达式为:

$$MSE(\hat{\theta})=E(\hat{\theta}-\theta)^2$$

(4)

其中, $\hat{\theta}$ 表示预测值, θ 表示真值。

3.3 猜拳游戏结果判断设计

猜拳游戏中有 3 种手势,石头、剪刀、布,石头赢剪刀,剪刀赢布,布赢石头。猜拳机器人和人的所有手势组合有 9 中,见表 1。交互结果较简单,使用决策树就能对游戏结果进行有效的判断。以机器人为决策节点,出拳动作为状态节点,最终结果为结果节点,决策过程如图 5 所示。

表 1 猜拳游戏胜负结果组合

人 机器人	石头	剪刀	布
石头	平局	机器人获胜	人获胜
剪刀	人获胜	平局	机器人获胜
布	机器人获胜	人获胜	平局

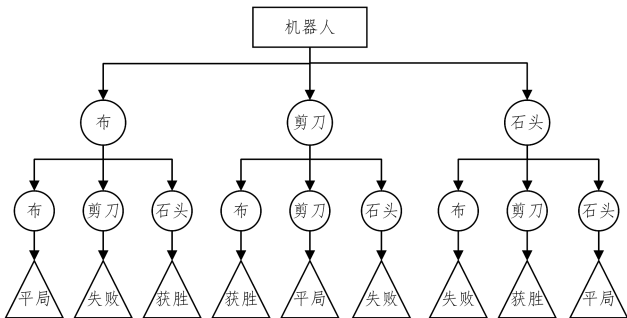


图 5 博弈结果决策树

3.4 系统实现

人机交互系统的流程如图 6 所示。NAO 机器人开机后进行系统初始化,连接上位机,启动视觉模块。准备就绪后,使用语音邀请人类玩家进行猜拳游戏,之后开始倒计时,在此过程中,调用预先训练好的 LSTM 人体行为预测模型对人类玩家的出拳意图进行预测。倒计时结束后,机器人根据预测结果出拳,人类玩家也已完成出拳动作。NAO 抓取人的手势图像,调用 ECPM 算法进行手势识别,识别结果与机器人自身的出拳一起以决策树进行博弈结果的判断。人类玩家每次的出拳手势会被记录在历史信息中,作为预测后续动作的数据来源。

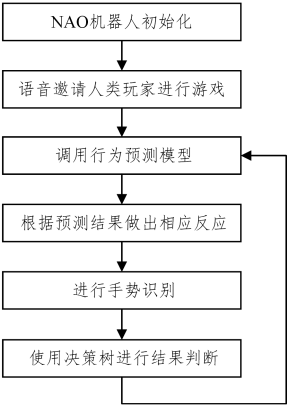


图 6 系统流程图

4 实验与结果分析

4.1 实验平台

本文设计的人机博弈系统以法国阿尔德巴兰公司生产的 NAO 仿人机器人为平台,采集图像数据传送至工作站。工作站对接收到的数据进行处理,发送控制指令至 NAO 仿人机器人,使其做出互动响应动作。

NAO 仿人机器人身高 57 cm,全身共有 25 个自由度,能够灵活地做出多种动作。配备 2 个高分辨率摄像头,最大分辨率为 1 280 × 960,支持无线网图像传输,具备语音输出功能,9 种语言,支持跨平台、多语言编程。

工作站硬件配置:E5-2620 v4 双 CPU,64 G 内存,500 G SSD,Quadro P5000 显示适配器,Ubuntu 18.04 操作系统。

4.2 实验结果与分析

4.2.1 手势识别结果分析

本实验采集 3 种手势不同尺度的关节特征热图各 1 000 张作为数据集,对 ECPM 网络进行训练。训练 10 个轮询后,在训练集上的准确率为 98.6%,在测试集上的准确率为 98.2%。之后再抓取 3 种不同手势不同尺度的图像进行测试,抓取到的图像如图 7 所示。

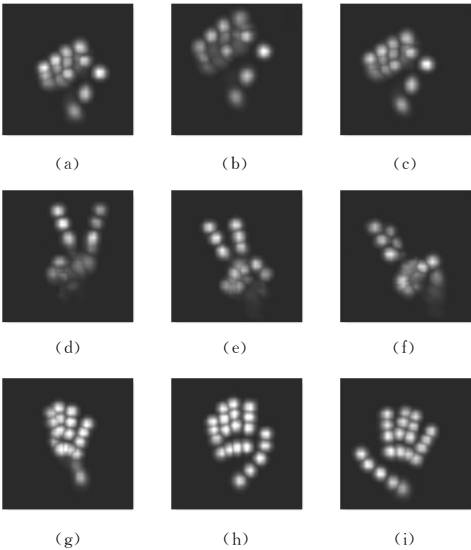


图 7 3 种不同手势不同尺度的关节特征热图

其中,图 7(a)一图 7(c)的识别结果为石头,图 7(d)一图 7(f)的识别结果为剪刀,图 7(g)一图 7(i)的识别结果为布。文献[7]使用传统的手势识别方法,识别过程为:1)手势图像

分割;2)图像预处理;3)模式匹配。使用 500 张手势图像进行测试,在简单背景下准确率最高为 92%,复杂背景下的准确率明显下降,最高为 83%。由于采用了模式匹配的方式进行识别,需为不同的手势预先设置特征模板,这导致不同手势的识别率不同。本文提出的方法首先利用姿态估计对手势的关节特征进行实时估计,再使用端对端的深度学习对关节特征热图进行识别并分类,克服了复杂背景下手势分割错误导致误识别的问题,并且手势类别、角度对识别结果无影响。为了进一步验证本文算法效果,将其与传统方法和几种其他深度学习方法^[14-15]进行比较,识别结果对比如表 2 所列。

表 2 识别结果对比

识别方法	平均识别率	
	手势类别	简单背景/% 复杂背景/%
文献[7]	石头	89 76
	剪刀	92 83
	布	92 80
文献[14]	96.5	
文献[15]	95.0	
本文方法	98.2	

文献[14]将手势图像转换为灰度图,通过阈值分割分离出手势部分,将其作为卷积神经网络的输入。文献[15]通过描边法在灰度图中提取手势边缘,再重建手势灰度图像,将其作为卷积神经网络的输入。

4.2.2 预测模块实验

由两名真人玩家进行猜拳游戏,记录其中一人的出拳顺序,共计 150 次。以 0,1,2 分别代表石头、剪刀、布,构建了一个时间序列{0,1,2,...}。将前 100 次的序列作为训练集,后 50 次的序列作为测试集,训练 epoch 设置为 100。预测结果的趋势图如图 8 所示。

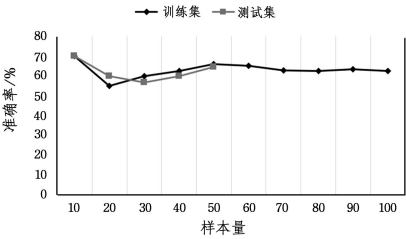


图 8 手势序列预测准确率曲线图

由图 8 可知,当数据量为 10 时,预测准确率达到 70%,随着数据量的上升,准确率稍有下降;在经过一个小幅度的波动后,准确率回升并趋于平稳。数据量超过 50 时,准确率稳定在 60%以上。最终在训练集上的预测准确率为 62.37%,测试集上的预测准确率为 64.44%,其结果表明有较好的稳定性。

结束语 本文提出了一种基于深度学习的手势识别算法,改善了传统基于图像分割——模式匹配方式的识别算法的进度。该算法结合姿态估计快速检测手势关节特征和卷积神经网络图像分类的优势,大大提高了手势识别的准确率,最终识别结果的准确率为 98%。文中设计了一个基于该算法的人机交互系统,以猜拳游戏的方式完成了基于视觉手势识别的人机交互过程。

参 考 文 献

[1] AGAH A. Human interactions with intelligent systems;Re-

search taxonomy[J]. Computers and Electrical Engineering, 2000,27(1):71-107.

[2] TADEUSIEWICZ R. Speech in human system interaction[C]// Human System Interactions. IEEE,2010:2-13.

[3] 齐静,徐坤,丁希仑. 机器人视觉手势交互技术研究进展[J]. 机器人,2017,39(4):565-584.

[4] PANWAR M. Hand gesture recognition based on shape parameters[C]//International Conference on Computing. 2012:317-319.

[5] RAUTARAY S S,AGRAWAL A. Vision based hand gesture recognition for human computer interaction;a survey[J]. Artificial Intelligence Review,2012,43(1):1-54.

[6] PISHARADY P K,SAERBECK M. Recent methods and databases in vision-based hand gesture recognition; A review[J]. Computer Vision and Image Understanding,2015,141:152-165.

[7] TSENG K T,HUANG W F,WU C H. Vision-based finger guessing game in human machine interaction[C]// 2006 IEEE International Conference on Robotics and Biomimetics. IEEE, 2006:619-624.

[8] LIN C Y,CHUANG L W,CHENG L C,et al. An interactive finger-gaming robot with real-time emotion feedback[C]// Proceedings of the 2015 6th International Conference on Automation. IEEE,2015:513-518.

[9] WEN X,NIU Y. A method for hand gesture recognition based on morphology and Fingertip-Angle[C]// 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE). 2010,1:688-691.

[10] 陈睿. 基于仿人机器人的博弈型人机交互研[D]. 上海:上海交通大学,2014.

[11] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016.

[12] HOCHREITER S,SCHMIDHUBER J. Long Short-Term Memory[J]. Neural Computation,1997,9(8):1735-1780.

[13] GRAVES A. Supervised Sequence Labeling with Recurrent Neural Networks[M]. Springer Berlin Heidelberg,2012.

[14] 蔡娟,蔡坚勇,廖晓东. 基于卷积神经网络的手势识别初探[J]. 计算机系统应用,2015,24(4):113-117.

[15] 王龙,刘辉,王彬. 结合肤色模型和卷积神经网络的手势识别方法[J]. 计算机工程与应用,2017,53(6):209-214.

(上接第 569 页)

数据加密、解密主要是对发送端的数据进行加密和对接受端收到的数据进行解密,系统采用 AES(高级加密标准)对数据进行加密,采用 128 位密钥,密钥由明文密码 MD5 形成 AES 加密密钥,这样就达到了双重加密的目的。只有收发双方都使用我们自行设计的数据采集终端,并且知道密码的情况下,才能实时接收所采集的数据,这就较好地保障了数据的安全性。

结束语 本文设计了基于 PXA270 的手持式数据采集终端,该终端集成了丰富的外设接口、有线网络通信和 WIFI、BT、GPRS 无线网络通信,能最大程度满足野外数据采集和传输需求。同时,基于 SIM300 模块的 AT 指令集,采用了 AES 和 MD5 双重加密技术,编制了加密的广域网移动通信软件,为野外条件下数据采集、传输提供了非常好的手段,具有广阔的应用前景。

参 考 文 献

[1] 田福平,田野,裴英博,等. 多样化军事任务异构数据集应用方法[J]. 火力与指挥控制,2015,40(7):164-167.

[2] 陈祥斌,张爱民,史芸. 通用装备保障数据资源建设关键问题研

究[J]. 军事交通学院学报,2014,16(11):22-26.

[3] 王鹏,田宗浩. 数据资源建设应着重解决的几个问题[J]. 中国西部科技,2015,14(1):81-82.

[4] 马海涛,关祥辰. 防空兵作战运用数据资源建设研究[J]. 防空兵学院学报,2014,31(1):23-25.

[5] 黄鹏良,陈钟荣,李杰,等. 一种便携式数据采集终端的设计与实现[J]. 信息化研究,2012,38(1):17-20.

[6] 勾国璐. 基于 S3C2440 和 Windows+CE 的信息交互终端硬件电路及驱动设计[D]. 南京:南京理工大学,2014.

[7] 邓志泉. 基于 Windows CE 操作系统的手持式物联网终端设计[D]. 成都:电子科技大学,2014.

[8] 李怀良,等. ARM 与 Windows CE 平台下小型化地震仪主机设计[J]. 科技通报,2016,32(10):121-143.

[9] 龚敏,方康玲,万鸣,等. 基于 EVC 的嵌入式导航电子地图设计[J]. 计算机应用,2009,29(10):2869-2874.

[10] 郑学敏,占红武,彭达,等. 基于 ARM-WINCE 系统的 EVC 图像处理界面开发[J]. 计算机应用与软件,2011,28(9):264-266.

[11] 杨梅,钟佩思,赵冠群,等. 基于嵌入式系统 Windows CE 的应用程序开发[J]. 制造技术与机床,2017,12(3):160-164.

[12] 解金耀. 基于 SMS 的无线信息采集系统的设计与设计[D]. 大连:大连海事大学,2005.