

基于 AHP 和生理电的机器人语音交互认知负荷评价

陈灏龙, 何人可

(湖南大学, 长沙 410082)

摘要: **目的** 提出一种兼具主观测量和生理数据验证的机器人语音交互认知负荷的评价方案。**方法** 设计实验使受试者对不同的机器人语音交互系统进行体验, 完成相应的任务并接受主观认知负荷测量, 同时加以生理电信号测量。主观测量部分利用层次分析法(AHP)对传统的 NASA-TLX 量表进行改进, 使得测量更加客观与细致; 并利用生理信号测量的结果验证主观评价的可靠程度与实际意义。**结果** 得出不同语音交互系统认知负荷的主观评价数据与生理测量数据, 两者相互印证, 从而得到可靠的认知负荷评价结果, 并进一步证明了评价系统本身的有效性。**结论** 基于 AHP 的基本原理, 改进了 NASA-TLX 量表, 并建立起了一套包含四个准则十四个指标的语音交互认知负荷评价体系, 利用心率与皮肤电信号测量进行了辅助验证, 提出了一种研究机器人语音交互用户体验的技术手段。

关键词: 语音交互; 认知负荷; AHP; 生理电; 机器人

中图分类号: TB472 **文献标识码:** A **文章编号:** 1001-3563(2020)18-0145-06

DOI: 10.19554/j.cnki.1001-3563.2020.18.023

Robot Voice Interaction Cognitive Load Evaluation Based on AHP and Physiological Signal

CHEN Hao-long, HE Ren-ke

(Hunan University, Changsha 410082, China)

ABSTRACT: It aims to introduce a method for evaluating the cognitive load of robot voice interaction based on both subjective measurement and physiological data verification. Experiments were designed for subjects to experience different robot voice interaction systems, finish different tasks and accept subjective cognitive load measurement and physiological electrical signal measurement. The analytic hierarchy process (AHP) was used in subjective measurement to improve the traditional NASA-TLX, which made the measurement more objective and detailed. And physiological signal measurement was used to verify the effectiveness and reliability of the subjective evaluation data. During experiments of different voice interaction systems, the subjective evaluation data and physiological measurement data of cognitive load were obtained. Reliable cognitive load evaluation results were obtained through their mutual verification, which further proved the effectiveness of the evaluation system. Based on the principles of AHP, the study improves NASA-TLX, establishes a voice interactive cognitive load evaluation system with 4 criteria and 14 indicators, completes auxiliary validation by heart rate data and skin electrical signal measurement, and proposes a technological means to research user experiment of robot voice interaction.

KEY WORDS: voice interaction; cognitive load; AHP; physiological signal; robot

机器人在人类日常生活中比重的上升, 使其与人之间的交互方式以及相关用户体验问题显得愈加重要。语音交互作为最基本的自然交互之一, 其用户体

验问题的研究方法尚未有定论, 本文通过引入认知负荷等心理学概念与层次分析法等数学手段, 并由生理电信号测量辅以验证, 力图构建一种适用于机器人语

收稿日期: 2020-05-06

作者简介: 陈灏龙(1994—), 男, 山西人, 湖南大学硕士生, 主攻信息与交互设计。

通信作者: 何人可(1958—), 男, 湖南人, 湖南大学教授, 主要研究方向为工业设计史、设计管理。

音交互行为的用户体验评价系统,对机器人语音交互过程中的认知负荷状况进行相对全面客观的评价。

1 机器人语音交互现状

机器人语音交互的研究一直处在学界前沿。2019年, Tsiourti 等人研究人们如何识别和响应类人机器人身体和声音所显示出的情绪^[1]。2018年, Du 等人研发出了一种语音与手势相结合的工业机器人在线教学方法,旨在提高人机协作教学效率^[2]。Martínez-Miranda 等人利用语音交互研究不同年龄段儿童对于机器人的情感反应和偏好^[3]。目前,此类研究多集中在技术层面,研究语音交互的用户体验问题,需要引入心理学概念。

2 认知负荷研究成果与应用

2.1 认知负荷理论

20世纪80年代末, Sweller 系统地论述了认知负荷理论,并定义认知负荷为“个体认知系统在一个特定作业时间内被施加的心理活动总量^[4]”。2017年, Krening 等人通过降低认知负荷的方式,提高机器辅助教学的效果^[5];2018年, Weber 等人尝试使用细粒度分析单元来解释认知负荷的发生^[6]。本文沿着这一思路,引入认知负荷的概念,并对传统测量方法加以改进,以期提出一种新的研究和评价思路。

2.2 认知负荷测量方法

主流的主观测量方法中, NASA-TLX 量表是由美国国家航天局提出的一种多维度工作负荷评估量表,要求受试者在完成一段任务后,通过比较心智需求、体力需求、时间需求、努力程度、受挫程度与自我绩效等六大指标权重,综合评估认知负荷的大小。2018年, Lee 等人以此探讨了智能手机形状因素对单手互动触摸行为的影响^[7]; Orlandi 等人则利用该量表测量了船舶操纵对海军飞行员造成的认知负荷^[8]。

由于高效、低成本,且可直接反映负荷大小,主观测量方法被广泛使用,但其结果的准确性较易受到受试者心理因素的影响,所以便须引入生理信号测量辅以验证。

2.3 生理信号测量

生理信号测量研究认知负荷,近年来发展迅猛。2017年, Mills 等人通过脑电图研究认知负荷^[9]; Abdelrahman 等人通过热成像仪的测量,研究了数字化系统对用户认知负荷大小的影响^[10]。2018年, Čegovnik 等人利用眼动仪测量驾驶过程中的认知负荷来评价驾驶的舒适度^[11]。

生理测量呈现客观性和实时性的特点,但由于实际影响因素复杂,使其结果常被夸大或削弱。与主观

测量同时进行并相互验证,可得到更具实际意义的结果。

3 基于 AHP 和生理电的研究方法改进

3.1 层次分析法

层次分析法(AHP)是美国运筹学家 Saaty 于20世纪70年代初提出的,是一种基于层次权重的决策分析法。2018年, Ergün 等人利用 AHP 确定了在机场保安服务中关键角色的相对重要性,使机场合理调配资源,从而提供最佳的安检服务^[12]; Haber 等人利用 AHP 对顾客需求进行分析,提出了一套产品服务系统开发流程^[13];2017年, Pan 等人则利用 AHP 的原理,提出了一套独立旅行计划的推荐系统^[14]。

由于可量化、简洁、易行等特点, AHP 及其原理在各领域广为应用。

3.2 评价方法体系的构建

NASA-TLX 量表须对各指标打分,并比较各权重指标,这样的操作流程和 AHP 的思路不谋而合,由此引入 AHP 对 NASA-TLX 量表进行改进,一方面使之更细致,更符合语音交互场景的测量要求,另一方面使 TLX 量表的权重评价更加清晰和准确,提高其本身的客观性与真实性。

跟据 NASA-TLX 量表与前期用户访谈,将 TLX 量表六大指标整理成四个主要评价准则和十三项评价指标。在准则层,心智需求(K_1)被分成了交互本身的复杂性(B_1)、交互方式学习与联想的难易程度,即交互迁移成本(B_2)、计算任务与记忆任务分别造成的负荷(B_3)、(B_4)以及交互系统本身的容错性,或者反过来说系统的脆弱性(B_5);体力需求(K_2)则被分为信息输入过程中视听上产生的疲劳(B_6)、输出过程中语言(B_7)和动手操作(B_8)两部分的疲劳,同时,由于传统 TLX 量表中的“时间需求(B_9)”与交互本身的时间节奏等属性息息相关,故拓展“自然流畅(B_{10})”与“反馈效率(B_{11})”指标,并统一在交互需求(K_3)准则之下;由于消费品使用中的“绩效体验”往往不甚明显,所以将传统 TLX 量表中有关用户主观体验的三个指标,即绩效(B_{12})、努力(B_{13})与受挫程度(B_{14})统一在主观体验(K_4)下,最终得出评价体系,机器人语音交互认知负荷评价体系见图1。

3.3 评价指标权重确定

据以上分析开展调研,共邀请三十六人,男生十七人,女生十九人,均为湖南大学师生,受试者专业不限,年龄相仿,且半年内无扫地机器人使用经历,初筛后共三十二人参加实验,男女各十六人。

设各指标 B_1, B_2, \dots, B_n 同处某准则 K 之下,用 Delphi 法征询三十二名受试者的意见,对同级不

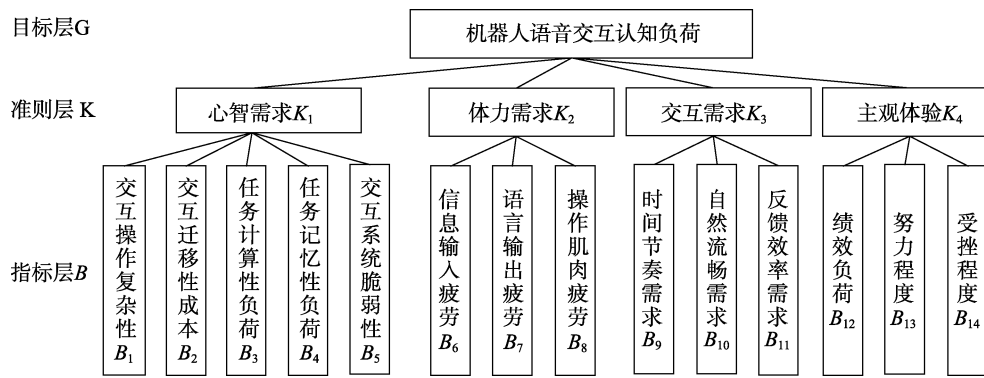


图 1 机器人语音交互认知负荷评价体系
Fig.1 Robot voice interaction cognitive load evaluation system

表 1 指标权重计算结果
Tab.1 Calculation result of index wght

准则层	对目标层权重	指标层	对准则层权重
K ₁ 心智需求	0.317567568	B ₁ 交互操作复杂性	0.262 295 082
		B ₂ 交互迁移性成本	0.191 256 831
		B ₃ 任务计算性负荷	0.147 540 984
		B ₄ 任务记忆性负荷	0.180 327 869
		B ₅ 交互系统脆弱性	0.218 579 235
K ₂ 体力需求	0.222972973	B ₆ 信息输入疲劳	0.346 153 846
		B ₇ 语言输出疲劳	0.41 025 641
		B ₈ 操作肌肉疲劳	0.243 589 744
K ₃ 交互需求	0.277 027 027	B ₉ 时间节奏需求	0.321 428 571
		B ₁₀ 自然流畅需求	0.401 785 714
		B ₁₁ 反馈效率需求	0.276 785 714
K ₄ 主观体验	0.182 432 432	B ₁₂ 绩效负荷	0.416 666 667
		B ₁₃ 努力程度	0.324 074 074
		B ₁₄ 受挫程度	0.259 259 259

同指标两两比较并打分,得各指标对准则 K 的重要程度。在记 b_{ij} 为 B_i 比 B_j 的情况下,如公式(1),可知判断矩阵 M_i :

$$M_i = \begin{bmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{bmatrix} \quad (1)$$

使判断矩阵 M_i 每一列归一化,如公式(2),并记矩阵 A 如公式(3):

$$a_{ij} = b_{ij} / \sum_{k=1}^n b_{kj} \quad i, j = 1, \dots, n \quad (2)$$

$$A = (a_{ij})_{m \times n} \quad (3)$$

正规化矩阵 A 的行和向量,排序权重向量 W 可得,如(4):

$$W = W_i / \sum_{j=1}^n W_j \quad (i, j = 1, 2, \dots, n) \quad (4)$$

指标权重计算结果见表 1。

一致性检验进行,随机一致性指标 (Random Consistency Index, RI) 取值见表 2,如式(5—7),分别求出最大特征根 λ_{max} 、一致性指标 (Consistency Index, CI) 与一致性比率 (Consistency Rate, CR),若计算可得 $CR < 0.1$,则各指标间良好一致性得以验证。

$$\lambda_{max} = \sum_{i=1}^n [(AW)_i / nW_i] \quad (5)$$

$$CI = \frac{\lambda_{max} - n}{n - 1} \quad (6)$$

$$CR = \frac{CI}{RI} \quad (7)$$

表2 随机一致性指标(RI)取值
Tab.2 Value of random consistency index

<i>N</i>	1	2	3	4	5	6	7	8	9
<i>RI</i>	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46

表3 表价系统一致性检验结果

Tab.3 Consistency check results of evaluation system

检验项目	λ_{max}	<i>CI</i>	<i>CR</i>
<i>G</i> 目标层	4.09	0.029 5	<i>RI</i> =0.89 (<i>n</i> =4) 0.033 2
<i>K</i> ₁ 心智需求	5.09	0.023 3	<i>RI</i> =1.12 (<i>n</i> =5) 0.020 8
<i>K</i> ₂ 体力需求	3.07	0.034 5	<i>RI</i> =0.52 (<i>n</i> =3) 0.066 4
<i>K</i> ₃ 交互需求	3.04	0.017 8	<i>RI</i> =0.52 (<i>n</i> =3) 0.034 3
<i>K</i> ₄ 主观体验	3.07	0.028 4	<i>RI</i> =0.52 (<i>n</i> =3) 0.054 7

评价系统一致性检验结果见表3,各层级 *CR* 均小于0.1,一致性均显示为“良好”,可以进行下一步实验。

4 机器人语音交互评价实验

4.1 实验受试者与设备

本实验环境为室内,噪声恒定50 Hz左右。BioLab多导生理仪为生理信号主要测量仪器;利用录音合成的语音交互系统三套,声音特质与操作指令(包括唤醒词)各不相同,某型号家用扫地机器人一台,并通过改装,增加蓝牙音箱模块,用于播放语音指令与反馈信息;具体过程有赖于摄像机和录音笔等设备进行记录;基于AHP的语音交互认知负荷主观测评表若干,以便给每段实验过程的认知负荷进行主观测量。

本实验受试者即上文指标权重确定环节的三十二名受试人员。

4.2 实验方法

预备时,生理信号贴片由实验人员为受试者黏贴,实验人员解释实验尝试解决的问题与实验过程中的动作规范,并引导受试者尝试进行语音交互。开始后,受试者依据任务脚本与不同交互系统,以特殊唤醒词等特定的语句对扫地机器人发出指令,完成操作流程;完成一次实验后,记录相应的心率与最大皮肤电信号数据(*GSR*_{max}),多次重复测量,心率取平均值,*GSR*_{max}取最大值。休息3 min后,更换语音交互系统,重新实验;实验结束,邀请用户在比较三次交互体验的基础上,填写主观测评表。

4.3 认知负荷主观评价

实验用认知负荷主观量表,由图1各项评价指标,依李克特(Likert)表形式制成,1分显示负荷很小,可忽略不计,5分显示负荷较高,让人明确感到不适。该问卷信度标准,即常用的克朗巴哈(Cronbach)系数 $\alpha=0.874>0.7$,落在“很可信”

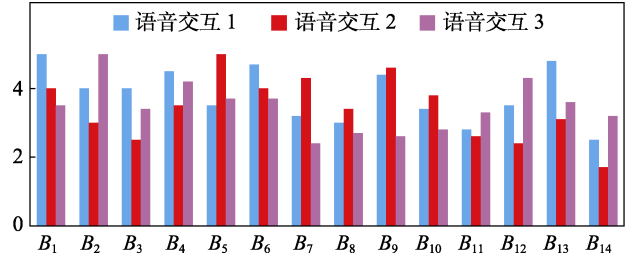


图2 认知负荷各项指标评价结果

Fig.2 Evaluation result of each cognitive load index

表4 基于AHP的认知负荷评价结果

Tab.4 Result of cognitive load evaluation based on AHP

评价项目	语音交互系统1	语音交互系统2	语音交互系统3
综合评价 (五分制)	1.677 127 658	1.561 709 019	1.761 163 322

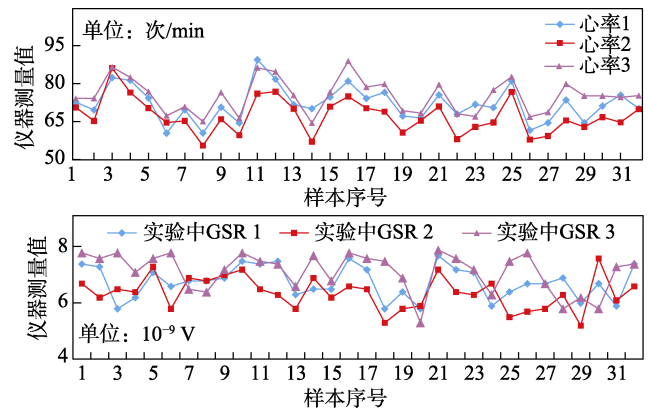


图3 实验生理电信号数据

Fig.3 Physiological electrical signal data during experiments

范围内,说明该量表有很高的可靠性、客观性与一致性。

三种语音交互方案各指标评价结果经过整理比对,认知负荷各项指标评价结果见图2,从中可以看到,系统1在*B*₁₋₄、*B*₆、*B*₁₂₋₁₄等项目中有较大的负荷得分,阐明该系统给用户较大的心智负荷与主观体验负荷,而系统2主要高负荷区则集中在在*B*₅、*B*₇₋₁₀等处,说明该交互系统有相对较大的交互需求负荷,三种语音交互系统中,系统3有最大的负荷得分,该系统在*B*₂₋₄、*B*₁₂₋₁₄、等区域都呈现出较大的负荷,与系统1呈现出相似的特征,而从系统3凸出的迁移性成本*B*₂与较大绩效负荷*B*₁₂可以得知,难于上手的操作方式与难以获得充足反馈的交互体验,也会对认知负荷最终的评价产生一定影响。见基于AHP的认知负荷评价结果见表4。

4.4 语音交互系统生理信号评价

实验过程中记录受试心率和最大皮电信号并按受试者序号绘成曲线,实验生理电信号数据见图3,并用SPSS 23.0软件对曲线上的样本数据进行LSD-t检验。

方差齐性检验见表5,方差齐性检验的显著性分

表 5 方差齐性检验
Tab.5 Homogeneity test of variance

	莱文统计	自由度 1	自由度 2	显著性
GSRmax	0.868	2	93	0.423
心率	0.022	2	93	0.978

表 6 单因素方差分析
Tab.6 One-Way ANOVA

	平方和	自由度	均方	F	显著性	
最大 GSR	组间	9.151	2	4.575	11.694	0.000 030
	组内	36.389	93	0.391		
	总计	45.540	95			
心率	组间	990.037	2	495.019	10.775	0.000 062
	组内	4272.668	93	45.943		
	总计	5262.705	95			

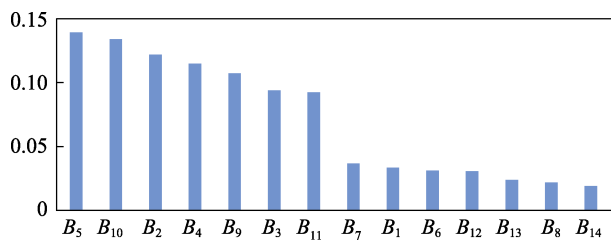


图 4 指标权重系统排序
Fig.4 Systematical ranking of index weights

别为 $P = 0.423 > 0.05$; $P = 0.978 > 0.05$ 说明方差数据有效, 继续进行单因素方差分 (One-Way ANOVA), 单因素方差分析见表 6, 显著性均小于 0.05, 说明数据间差异性显示为“非常显著”。

进行事后 LSD-t 检验, 两两配对探索数据间相关关系, 得到最大 GSR 测量值之间差异显著, LSD-t

检验见表 7, 显著性均小于 0.05, 并且根据平均值差值亦可以得到与 AHP 评价相似的认知负荷大小排序; 而心率数据 LSD-t 检验结果则表明系统 1 和系统 3 之间的数据差异性不够明显 $P = 0.081 > 0.05$, 一方面印证了主观分析结果中, 系统 1 和 3 之间的相似性, 另一方面, 也有可能因为实验进行时间较长, 使得受试者心理状态产生波动, 影响数据质量。总的来说, 基于 AHP 的主观认知负荷评价得到了来自生理信号结果的验证。

5 实验结论

本文基于 AHP 对传统 NASA-TLX 量表进行了改进, 提高了 TLX 量表的准确性, 并据 NASA-TLX 原理和语音交互特性建立评价体系。对评价指标再次对目标层进行排序后, 指标权重系统排序见图 4。

显然, 评价指标中, 有一类更为重要, 如系统的脆弱性 (B_5)、自然流畅性 (B_{10})、交互迁移成本 (B_2)、记忆任务造成的负荷 (B_4)、时间需求 (B_9)、计算符合 (B_3) 以及语音交互系统的反馈效率 (B_{11}) 等, 称之为主要指标, 剩下的则称之为次要指标。

主要指标涵盖心智需求和交互需求两大准则层, 因为实验本身的体力脑力消耗并不大, 所以体力需求准则下的各个指标相对不是很重要; 而自我绩效准则也属于次要指标范围之内, 可能源于项目挑战性不够强, 亦可能在受试者眼中, 此类指标较难描述或比较, 从而造成系统误差。

对于设计者来说, 一方面要提高整体的体验流畅度, 降低用户的记忆与计算任务强度, 保证语音交互系统的抗干扰能力, 另一方面, 提高系统本身的通用性与易用性也是降低认知负荷的重要手段。

表 7 LSD-t 检验
Tab.7 LSD-t test

因变量	(I) 组别	(J) 组别	平均值差值 (I-J)	标准误差	显著性	95%置信区间	
						下限	上限
最大 GSR	1.00	2.00	0.37500*	0.15638	0.018	0.0645	0.6855
		3.00	-0.38125*	0.15638	0.017	-0.6918	-0.0707
	2.00	1.00	-0.37500*	0.15638	0.018	-0.6855	-0.0645
		3.00	-0.75625*	0.15638	0.000	-1.0668	-0.4457
	3.00	1.00	0.38125*	0.15638	0.017	0.0707	0.6918
		2.00	0.75625*	0.15638	0.000	0.4457	1.0668
心率	1.00	2.00	4.80469*	1.69453	0.006	1.4397	8.1697
		3.00	-2.99156	1.69453	0.081	-6.3566	0.3734
	2.00	1.00	-4.80469*	1.69453	0.006	-8.1697	-1.4397
		3.00	-7.79625*	1.69453	0.000	-11.1612	-4.4313
	3.00	1.00	2.99156	1.69453	0.081	-0.3734	6.3566
		2.00	7.79625*	1.69453	0.000	4.4313	11.1612

注: *为平均值差值的显著性水平为 0.05

6 结语

近年来,对机器人语音交互系统的研究多集中在语音识别与系统构建等硬件开发层面,较少讨论其人机交互的体验问题。本研究设计实验,对三种机器人语音交互系统进行评价,利用AHP改进NASA-TLX量表,并利用Bio-lab生理信号仪监测用户的心率变化和皮电反应,进一步印证了主观测量的结果,从而验证了整个认知负荷评价体系的可靠性与准确性。同时,通过评价体系内部不同评价指标之间的比较,得到语音交互系统不同特征对整体认知负荷的贡献,更好地帮助设计者找到机器人语音交互系统的问题所在。

在取得了较为理想的数据表现的同时,本研究也有很多不足,如受试者均为湖南大学师生,可能会对结果造成偏差;另外从数据表现,猜测生理测量结果的波动也可能来源于实验时间持续导致的疲劳和焦躁,但数据波动并未对最终结果产生较大影响,待后续研究进行甄别。从整体来讲,本文为研究机器人语音交互过程的认知负荷与用户体验问题,提出了一种比较可靠的技术路径。

参考文献:

- [1] TSIOURTI C, WEISS A, WAC K, et al. Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In) Congruence on Emotion Recognition and Attitudes Towards Robots[J]. *International Journal of Social Robotics*, 2019, 11(4): 555-573.
- [2] DU G, CHEN M, LIU C, et al. Online Robot Teaching with Natural Human-Robot Interaction[J]. *IEEE Transactions on Industrial Electronics*, 2018, (99): 1-1.
- [3] MARTÍNEZ-MIRANDA J, PÉREZ-ESPINOSA H, ESPINOSA-CURIEL I, et al. Age-based Differences in Preferences and Affective Reactions Towards a Robot's Personality during Interaction[J]. *Computers in Human Behavior*, 2018, (84): 245-257.
- [4] SWELLER J. Cognitive Load during Problem Solving: Effects on Learning[J]. *Cognitive Science*, 1988, 12(2): 257-285.
- [5] KRENING S, HARRISON B, FEIGH K M, et al. Learning from Explanations Using Sentiment and Advice in RL[J]. *IEEE Transactions on Cognitive and Developmental Systems*, 2017, 9(1): 44-55.
- [6] WEBER B, NEURAUTER M, BURATTIN A, et al. Measuring and Explaining Cognitive Load during Design Activities: a Fine-Grained Approach[J]. 2018.
- [7] LEE S C, CHA M C, HWANGBO H, et al. Smartphone form Factors: Effects of Width and Bottom Bezel on Touch Performance, Workload, and Physical Demand[J]. *Applied Ergonomics*, 2018, (67): 142-150.
- [8] ORLANDI L, BROOKS B. Measuring Mental Workload and Physiological Reactions in Marine Pilots: Building Bridges Towards Redlines of Performance[J]. *Applied Ergonomics*, 2018, 69: 74-92.
- [9] MILLS C, FRIDMAN I, SOUSSOU W, et al. Put Your Thinking Cap on: Detecting Cognitive Load Using EEG during Learning[C]. *The Seventh International Learning Analytics & Knowledge Conference*. ACM, 2017.
- [10] ABDELRAHMAN Y, VELLOSO E, DINGLER T, et al. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load[J]. *Proceedings of the Acm on Interactive Mobile Wearable & Ubiquitous Technologies*, 2017, 1(3): 33.
- [11] ČEGOVNIK T, STOJMENOVA K, JAKUS G, et al. An Analysis of the Suitability of a Low-cost Eye Tracker for Assessing the Cognitive Load of Drivers[J]. *Applied Ergonomics*, 2018, (68): 1.
- [12] ERGÜN N, BÜLBÜL K G. An Assessment of Factors Affecting Airport Security Services: an AHP Approach and Case in Turkey[J]. *Security Journal*, 2018, (1): 1-25.
- [13] HABER N, FARGNOLI M, SAKAO T. Integrating QFD for Product-service Systems with the Kano Model and Fuzzy AHP[J]. *Total Quality Management & Business Excellence*, 2018, (9): 1-26.
- [14] PAN Q, WANG X. Independent Travel Recommendation Algorithm Based on Analytical Hierarchy Process and Simulated Annealing for Professional Tourist[J]. *Applied Intelligence*, 2017, (4).