A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance

Eoghan Hynes[1], Ronan Flynn[1], Brian Lee[2], Niall Murray[1]

Department of Computer and Software Engineering[1]
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
email: {e.hynes, nmurray}@research.ait.ie[1], rflynn@ait.ie[1]

Software Research Institute[2]
Athlone Institute of Technology
Athlone, Co. Westmeath, Ireland
email: blee@ait.ie[2]

***Abstract*— Augmented reality (AR) can support a user in performing an expert task by overlaying real world objects with the domain specific information required to complete the task. Understanding how users can process and use such information is very important for informing the design of AR technologies and applications. In this paper, the results of a quality of experience (QoE) evaluation of an AR application for the task of solving a Rubik's Cube are presented. The Rubik's Cube was selected based on its familiarity and the expertise needed to solve it unaided. An empirical approach was taken to identify the QoE features that affect the usability and utility of an AR head-mounted display (HMD) compared with paper-based instruction. The QoE evaluation methodology involved the capture and analysis of implicit and explicit QoE metrics. The utility (in terms of performance) of each mode of instruction was objectively measured using: (a) cube completion success rates; and (b) time-to-completion. The implicit metrics of electrodermal activity (EDA), skin temperature, heart rate and the novel use of facial action units (AUs) were recorded to infer emotional state during the task completion. Finally, with respect to explicit metrics, the test subjects completed a Likert scale questionnaire post the experience to subjectively report QoE as well as a self-assessment manikin (SAM) questionnaire to self-report emotional state upon task completion. The results show that AR yielded higher success rates and significantly lower time-to-completion rates. The AR group explicitly reported higher levels of positive valance (affective state) than the paper-based group. The physiological data showed that the AR group were less stressed (via EDA) than the paper-based group. Finally, analysis of the AU data reflected a greater than chance (total: 21.85%) accuracy when predicting affective state based on SAM questionnaires as ground-truth.**

Keywords*— *quality of experience, augmented reality, facial action units, self-assessment manakin, arousal, valence.

I. INTRODUCTION

It is crucial that systems that aid users in complex task assistance are evaluated with respect to their usability and utility to support the effective, efficient and satisfying experiences for users [1]. As technology becomes more ubiquitous and embedded in workplaces, users will be required to perform tasks with greater degrees of complexity and customisability [2]. For example, an emerging use case for AR is in mass customised assembly assistance in Industry 4.0 [2].

With the recent emergence of AR based HMDs, there has been significant interest with respect to AR based task assistance across a range of application domains for improved productivity and efficiency [2], [3]. However, there are a lack of studies that consider the user as a key stakeholder in this process and indeed their perception of using AR as a task assistance tool. Key elements of user performance such as task duration, success rates, ease of interaction and usability are critical factors to consider. Recent works have acknowledged the importance of understanding these aspects in domains such as Industry 4.0 [3].

One approach to understanding these elements is through analysing the factors that influence the user's perceived QoE. The definition of QoE includes "the degree of delight or annoyance of a person whose experiencing involves an application, service, or system" [1]. Research has shown that user QoE can be influenced by several factors, including technical, social, and psychological. Indeed QoE evaluation methodologies are still an active research topic [1]. The traditional approach has been through the use of explicit evaluations i.e. the use of post experience questionnaires. However, research has highlighted issues with Mean Opinion Score (MOS) based approaches [4] and more recently, the research community has begun to investigate the value of metrics such has EDA and Heart Rate during the experience [5]–[7].

In this work, we present the design and evaluation of an AR based Rubiks cube solving application and compare it with paper-based instructions. The Rubik's Cube solving task was selected due to its complexity (i.e. standard 3x3 Rubik's Cube has 43 quintillion possible states - an NP-complete problem [8]) and its familiarity. For each instruction modality, we employed implicit and explicit data capture and analysis. The results highlight potential for AR with respect to productivity however, they also highlight issues with AR with respect to aesthetics. This is important because it informs the user centred design of AR based technologies.

The rest of this paper is structured as follows: section II presents research works that have performed evaluations of AR; section III outlines the task instruction presentation approaches (i.e. AR and paper-based instruction); section IV presents the explicit and implicit data capture approaches; section V presents the experimental method; section VI presents the results and a discussion of the results, whilst section VII presents the conclusion.

II. RELATED WORK

This section highlights research works that have compared AR with other modalities for task-based instruction applications. In particular, it highlights how our work addresses a key issue, a lack of consideration for the user.

In [9], a visual analytics application named HoloBee, (originally presented in [10]) for visualising the phenomenon of bee drift was presented. The application was presented on the Microsoft™ Hololens™ and a desktop-based interface. A within-subjects comparative QoE evaluation was conducted for each interface to evaluate task performance and user experience (UX). Test subjects performed 3 information analysis tasks of varying complexity on the bee drift data using each interface. Each test subject completed a 7-point Likert scale questionnaire after using the application on both interfaces. The questionnaire considered the user UX criteria: (i) intuition; (ii) ease of use; (iii) comfort; (iv) naturalness and (v) efficiency of each of the display interfaces. The goal of this study was to identify the impact of UX criteria on interface

978-1-7281-1817-8/19/$31.00 ©2019 European Union

preference. Although the authors did not report which interface was preferred overall, the results showed that for all considered aspects (i-v), efficiency was the UX criterion that correlated strongest with interface preference. Comfort was the criterion that correlated weakest with interface preference.

In [3], the authors compared AR and Video for assembly assistance using Lego™. Each mode of assistance was assessed objectively for accuracy (task completion (TC)) and performance (time-to-completion (TTC)). The focus of this research was to determine the likelihood of user adoption of AR for assembly assistance. To this end, the authors also partially employed the Technology Acceptance Model 3 (TAM3) [11]. Within the 'use intention' metric they selected the 'perceived ease of use' determinant which consists of six sub-determinants. From these six, they used (a) 'perceived external control' and (b) 'perceived enjoyment'. In addition, the authors sought to evaluate the mental workload of completing the Lego™ assembly task using the NASA-TLX. The results showed a reduction in TTC and 'mental' workload, and a statistically significant reduction in the number of errors for AR instruction. The authors reported, via multiple regression analysis, that 'perceived enjoyment' has a significant effect on 'use intention'. The insignificant effect of 'perceived external control' was disregarded considering their testing environment.

AR was compared with written documentation as a guidance tool for a Lego™ assembly task in [2]. The paper outlined two experiments. The first experiment used the NASA Task Load Index (TLX) [12] to assess total task load between AR and user-manual based guidance. The second experiment assessed the learning curves between the two guidance modes. The results showed: shorter task completion time; lower total task load; fewer assembly errors; and a lower learning curve for the AR group. The authors highlighted issues with respect equipment (sensor) complexity and technical constraints for the evaluation system.

Closely related to the work presented here, the authors of [13] & [14] outlined an approach to testing the utility of AR as an assistive technology for a Rubik's Cube solving task. In [13], an AR application was developed in Unity™ using the Vuforia [1] AR Software Development Kit (SDK). This application used the Kociemba algorithm to calculate an optimal cube solution. A 3D Rubik's Cube was rendered and solved on-screen when an AR target was detected. The instructions for solving the cube were rendered on screen. However, the AR application did not verify successful step completion and instructions were rendered based on paper template matching (i.e. it didn't scan a real-world cube to determine cube state as in our work). In [14], an AR application, also developed in Unity™ using the Vuforia AR SDK, supported the solving of a real world 2x2 Rubik's Cube. This experiment used a desk-mounted iPad 2™ to execute the AR application. The larger squares on a 2x2 cube facilitated cube detection for augmentation in Vuforia. The developers also placed markers (fiducials) on the cube squares to enable detection of the cube. A pattern database was used to display the cube solution to the user on the iPad screen.

Considering the related works, this work contributes to the state of the art through the novel capture of explicit and implicit metrics for task assistance-based AR. More specifically our work involves real-time recording of performance, physiological and facial action units and post-test recording of Likert and SAM questionnaire responses. Therefore, this holistic QoE evaluation of AR is undertaken to determine the factors that affect acceptability of AR for complex task assistance.

III. Immersive and Paper based Presentation Systems

In this section, the AR and paper-based assistance modalities are described.

A. Augmented Reality Presentation System

The AR group instructions were presented using the Meta2[2] AR HMD. The Meta2 was a prototype AR HMD aimed at developers where augmentations are rendered on a screen positioned in front of the wearer's eyes. The Meta2 boasted a 90-degree field of view, 2.5K screen resolution with a 60Hz refresh rate, a 720p front-facing RGB camera and 9 ft (2.7 m) USB cable for video, data & power.

Unlike the Rubik's Cube based tasks in [13] and [14], image-based template matching was not used in the development of the AR application in this research. The Vuforia AR SDK was found to be incapable of standard marker-less Rubik's Cube detection. Vuforia uses image-based template matching for object registration and tracking. The more intricate the patterns in the template image, the better the tracking. The inability to detect objects which lack intricate patterns is a perceived weakness of the image-based template matching approach to AR development. AR developers often resort to labelling objects with fiducials (QR codes, bar codes) to allow for object detection using the image-based template matching approach to AR. Instead, C# was used to hard code a geometric description of a standard 3x3 Rubik's Cube. OpenCV filters were then used for real time detection of this pattern in the video feed. Using OpenCV filters and software to encode and detect a geometrical description of the cube's colours and contours in the video feed was a lower level approach than image-based template matching, and allowed for registration and tracking of the fiducial-less Rubik's Cube. The AR application was adapted from an online repository [3], originally developed for Android™[4] devices. This application was translated from Java into C# using a C# wrapper for the C++ OpenCV library in Unity™ for use with the Meta2 HMD. This application can be readily ported from Unity™ to many other AR platforms.

The AR application used the Kociemba algorithm [15] which solves the cube with the least possible number of moves (optimally) from any scrambled state. At the beginning of each test, the front-facing camera on the Meta2 was first used to scan all faces of the scrambled cube. It detected the state of the scrambled Rubik's Cube using a combination of OpenCV filters, (Y,U,V) colour and shape detection algorithms. Once the Cube was successfully scanned, the AR application then proceeded to heuristically step through a two-layer deep combination of moves towards the solved state. It selected the shortest solution at each step, typically arriving at the optimal solution for solving the Rubik's Cube without need for resource intensive verification passes. The steps to solve the cube were displayed in the user's field of view in sequence. The instruction consisted solely of a line of text, describing

[1] https://developer.vuforia.com/
[2] https://metavision.com/
[3] https://github.com/AndroidSteve/Rubik-Cube-Wizard
[4] **Android** is a **trademark** of Google LLC.

the angle and direction to turn the face in question. In Rubik's Cube nomenclature, a face is referenced by the tile at its centre, because each centre tile is bound to one face. The face names in the instruction were colour coded. An example of such an instruction is:

"Rotate the face with the Red tile at its centre 90º clockwise."

B. Paper Based Presentation System

The comparison group was provided with the same set of instructions as the AR group, but printed on paper. The paper-based instructions were presented in a 22-page A4 manual. Each page consisted of one text instruction, where the name of the face was coloured the same as in the AR instructions.

IV. Explicit and Implicit QoE Instruments

QoE was explicitly reported using Likert and SAM questionnaires. The Likert scale questionnaire was designed to build a metric of QoE consisting of fourteen questions. Each question mapped to at least one of the six aspects of QoE identified in [16]. The six determinants of QoE are utility, usability, interaction, aesthetics, efficiency and acceptability. Five questions mapped to interaction, four to usability, two to utility and one each to aesthetics, acceptability and efficiency QoE criteria.

Affective state was explicitly reported using the SAM questionnaire. The SAM questionnaire consists of three scales, one for each dimension of affect (arousal, valence and dominance). The test subjects completed the questionnaire by circling one manakin on each scale representing the level of the dimension that they felt upon task completion. Valence can be defined as "the dimension of experience that refers to hedonic note. Arousal describes the level of energy in the hedonic note" [17]. Dominance relates to the feeling of submissiveness to the stimulus. It accounts for the least amount of variance in affective judgements [18].

EDA, skin temperature and heart rate were recorded using the Empatica E4 wristband device [19] to implicitly evaluate arousal. A five-minute baseline reading was taken from each subject prior to testing. The baseline data provided the standard physiological reading for each individual under test, from which any in-test deviations were calculated in time-series analysis. Positive deviations are reported herein to be indicative of heightened arousal as an objective metric of QoE. This was the approach taken in [20] for EDA. Heightened arousal is considered herein in combination with the positive or negative valence component ('*delight*' or '*annoyance*').

The emotional signal that is encoded and transmitted in facial expressions lends itself well to the implicit evaluation of user QoE under its definition as "The degree of delight or annoyance…". As such, test subject AUs were recorded using a desk mounted Logitech 1080p video camera and the '*OpenFace*'[5] facial recognition application. The open source software was edited to classify emotions based on the presence of combinations of AUs based on the rules outlined in [21] and [22]. Emotions were classified into one of five groups. One for each quarter of the 2D circumplex model [17] and one for neutral. These were *Delighted*, *Sad*, *Content*, *Annoyed* and *neutral*. Only lower facial action units were used because the AR group's upper facial AUs were occluded by the AR HMD. The majority of facial expression of emotion occurs in the lower half of the face [21]. The 2D Circumplex model was then used to decompose the AU based emotion into its component dimensions inspired by the work of [17]. In this way, AUs were compared to SAM responses for multidimensional accuracy in terms of polarity (not intensity). These implicit QoE metrics were recorded continuously throughout the baseline training, practice and testing phases.

V. Methodology

The experimental method outlined in this section includes six key phases: Sampling, Screening; Baseline; Training; Practice; and Testing. The methodology used extends those outlined in section II by recording the test subjects physiological, facial AUs, and self-reported emotional state. The methodology adheres to the standards for subjective and objective quality evaluation outlined in ITU-T P.913[23].

A. Sampling and Information Sharing Phase

A convenience sampling approach resulted in 48 individuals taking part in experiment with a mean age of 32 years (standard deviation (SD): 10 years). Equal gender representation was sought in the sample group in line with clause 9.3 of ITU-T P.913 [23]. The 48 participants were divided into two groups of 24 consisting of 12 males and 12 females in each group (human factors have been shown to influence QoE [24]). Each participant was provided with a test information sheet explaining the experiment in full. After reading this, test subjects completed a consent form.

B. Screening Phase

In the screening phase, test subjects were first screened for visual acuity using the Snellen test and then screened for colour perception using a digital [6] Ishihara-based test. Following this, an interactive digital [7] Vandenberg-based mental rotation test was implemented. This test provides a direct and convenient measurement for human capacity of spatial cognition [2]. No test subjects were excluded during screening.

C. Baseline Phase

The test subjects were seated at a table in a controlled lab environment. To record the subjects baseline EDA, heart rate and skin temperature ratings, they were fitted with an Empatica E4 wrist band device. A five-minute reading of this physiological data was undertaken prior to the training phase. This data was used to calculate an individual's mean EDA, skin temperature and heart rate at rest. Any in-test deviations from these individual standard readings would later be calculated as a measure of in-test arousal levels. This in turn is combined with other explicit and implicit affective state data and considered as a metric to evaluate QoE. At this point, a desk mounted video camera began recording head pose data using the '*OpenFace'* facial recognition software. This data included facial AUs and continued throughout the training and testing phase.

D. Training Phase

As part of the methodology, all test subjects underwent a training phase. Written instructions were provided that included Rubik's Cube solving terminology. The instructions outlined a description of the mode of instruction they would be using (paper-based or AR) and what they should do to

[5] https://github.com/TadasBaltrusaitis/OpenFace
[6] http://www.color-blindness.com/ishihara_cvd_test/ishihara_cvd_test.html
[7] http://vample.com/tools/mental-rotation

complete the test. There were three fundamental instructions used to solve the Rubik's Cube. These were: (a) a 90º clockwise, (b) a 90º anti-clockwise, or (c) a 180º clockwise rotation, of the face in question. The face in question is identified by referring to the colour of its central tile, as this never changes location on the face. The training phase also verified that participants understood the Rubik's Cube manipulation terminology. In the training phase, the test subjects were provided with a randomly scrambled Rubik's Cube and asked to manipulate the cube in the manner described by the instructions. The average training phase took under one minute for both groups. After verifying that the test subjects understood the instructions, the practice phase began.

E. Practice Phase

This phase consisted of a potential maximum of three practice runs. Each practice run included a fixed set of six instructions. Each set of instructions consisted of two of each of the three types of manipulation, one for each face of the cube. The six colours of a standard 3x3 Rubik's Cube are blue, green, white, yellow, orange and red. The AR group was introduced to a randomly shuffled Rubik's Cube and fitted with the Meta2 AR HMD. A personal computer keyboard was positioned at the centre of the table. The AR group progressed through the AR application by pressing the space bar on the keyboard to deliver each instruction in turn. The control group (CG) was presented with a randomly shuffled Rubik's Cube and an instruction manual containing the same set of six instructions as the AR group, with one instruction per page. The CG progressed through each instruction by turning each page of the instruction manual in turn. The assessor recorded the turning of each page by pressing the space-bar key on a keyboard. Test subjects attempted to follow each instruction in turn by manipulating the Rubik's Cube as instructed. In the training phase, the correct following of all instructions did not end with a correctly solved Rubik's Cube. The space-bar presses were recorded to reflect instruction intervals for both groups. Practice run durations, number of required practice runs, total errors and the index of incorrectly followed instructions were recorded to assess the learning curve of both the AR and written modes of instruction delivery. If a test subject made an erroneous cube manipulation, they were afforded a further practice run. The maximum number of practice runs required by either group to follow all instructions successfully was two runs. When the training phase ended the test subject proceeded to the testing phase.

F. Testing Phase

In the testing phase, the test subjects were presented with the test Rubik's Cube in the superflip position. The superflip position has the furthest distance from the solved state and takes a full suite of 20 moves to solve using the AR application. The starting point of the superflip position facilitated commonality across all subjects. The AR group proceeded to the testing phase wearing the AR HMD while the CG was presented with the test instruction manual. If the test subjects followed each step correctly, they ended the test with a correctly solved Rubik's Cube.

Successful cube completion; Time-to-completion; number of errors and the index of incorrectly followed instructions were recorded during the testing phase. The test ended after recording the final (21st) instruction, which simply stated that 'The cube should now be solved'. After the test, the test subjects completed the five-point Likert scale questionnaire reporting on their subjective experiences during the test in terms of interaction, efficiency, usability, aesthetics, utility and acceptability as outlined in [16]. They also completed the nine-point SAM questionnaire to evaluate their emotional state at test-end, inspired by the work of [25].

VI. Results and Discussion

In this section of the paper the results and the analysis of the implicit and explicit metrics as part of the QoE evaluation are presented.

A. Implicit Metrics: TTC, TC, Physiological, Facial AUs

Results show 91.67% and 95.83% successful task completion for the CG and AR groups respectively. An independent samples T-test using SPSS [8] showed that the inter-group success rates were not statistically significant with p=0.555. The mean task completion time was 142.21 seconds for the AR group and 162.29 seconds for the CG. The inter-group test duration difference was statistically significant with p=0.040 at 95% confidence. These findings are in line with [3] , [3] and show that AR offers efficiency and productivity gains.

Because the physiological data were not normally distributed, a non-parametric Mann-Whitney U-test was undertaken during statistical analysis. Physiological results showed that the CG had higher deviations (baseline -> test) of skin temperature (+ 1.4ºC, p=.273) , heart rate (+ 0.8 BPM, p=.613) and significantly higher deviations of EDA (+ 0.56 µS, p=.000). This indicates that the CG experienced higher in-test arousal levels than the AR group. These results are shown in Table I. An investigation into the nature of this heightened arousal is presented in the next section.

Regarding emotional state classification using AUs, the AR group showed higher classifications of both *Delight* and *Annoyance* and lower classifications of *Sad* and *Neutral.* The multidimensional components of these results are shown in Table II. The inter-group difference for AUs was not statistically significant with a probability value of p=0.266 at 95% confidence.

B. Explicit Metrics: SAM & Likert Questionnaires

The SAM questionnaire asked the test subjects to recall how they felt at the moment of task completion. The polarity (not intensity) of SAM questionnaire results is seen in Table II. Table II shows that combined positive valence and positive arousal were reported in 62% and 87% of instances for the CG and AR group respectively. The inter-group difference of SAM results was statistically significant with p=.033, with the main distinction lying between levels of *Delight* (+V +A) and C*ontentment* (+V -A). The SAM data shows that the heightened arousal state of the CG seen in the physiological data is 28.57% less *Delight* related than for the AR group.

A comparison of AU based valence and arousal estimation against the SAM results as ground truth was undertaken. The results of this comparison are seen in Table III. For this comparison, the mean classification from a 6 second window of AU classifications was used. This window size is prescribed in [26], allowing for onset-apex-offset of the expression. The accuracy (TP+TN+T0/Total) of combined valence and arousal estimation is 22.9% and 20.8% for the CG and AR group respectively. Hence, lower face AUs were 6.2%

[8] https://www.ibm.com/analytics/spss-statistics-software

TABLE I. STATISTICAL ANALYSIS RESULTS OF PHYSIOLOGICAL DATA (95% CONFIDENCE LEVEL)

	AR mean	CG mean	Sig. (2 tailed)
EDA (μS)	-0.29	0.27	.000
BPM	0.65	1.45	.613
ST[a] (ºC)	0.58	0.72	.273

a.ST: Skin Temperature, Sig: Statistical Significance.

TABLE II. QUANTITIES OF VALENCE AND AROUSAL COMBINATIONS

Type	Group	+V +A[a]	+V -A	-V +A	-V -A	Neut
AU	AR	2	0	6	13	3
	CG	1	0	4	14	5
SAM	AR	21	2	0	0	1
	CG	15	7	0	1	1

a. V: Valence, A: Arousal, Neut: Neutral.

TABLE III. AU BASED ESTIMATION OF VALENCE AND AROUSAL COMPARED TO SAM RESULTS

Group	Type	TP[a]	TN	FP	FN	T0	F0	Acc.
AR	Valence	2	0	0	19	0	3	20.8%
	Arousal	7	1	1	12	0	3	
CG	Valence	1	0	0	18	0	5	22.9%
	Arousal	4	5	1	9	1	4	

[a].TP: True Positive, FN: False Negative, F0: False Neutral, Acc: Accuracy.

and 4.1% more accurate than chance (1/6 = 17%) at estimating valence and arousal combinations when compared to the self-reported SAM results for the CG and AR groups respectively. [21] notes that the majority of AUs are more effectively used for the expression of negative emotions, which can be seen in Table II. The difference between negative valence in AU results is betrayed by the same amount of positive valence explicitly reported in the SAM results for both groups (i.e. 4).

Table IV shows the Likert questionnaire questions, including their mapping to QoE aspects, the MOS for both groups and the statistical significance of differences between the groups. A non-parametric Mann-Whitney U-test was undertaken to analyse the questionnaire results for statistical significance. These results showed that the groups differed significantly on three questions. These were question 1, 3 and 8. Question 1 mapped to utility in terms of usefulness of the instructions. It can be seen in Table IV that the MOS for interaction was higher for the CG at 27.00, than for the AR group at only 22.00. For this evaluation, the instructions consisted of text only to ensure equal modalities for both groups. This may be considered a limitation of this evaluation. AR test subjects may have legitimately anticipated more interactive instructions, which is reflected in this result. Question 3 was mapped to the aesthetics aspect of QoE under comfort. Table IV shows that the MOS of question 3 was higher for the AR group at 29.42 compared to the CG at 19.58. As this question had a negative connotation, it signifies that the AR group reported experiencing higher discomfort. The difference in reported comfort comes from bespectacled test subjects reporting less comfort with the HMD in the AR group, where 6 and 5 test subjects wore glasses in the CG & AR group respectively. The Meta2 AR HMD is designed for use with spectacles but causes some pressure at the side of the head. This left a temporarily visible mark on some bespectacled test subjects after use. Question 8 mapped to interaction quality in terms of the instructions being distracting. The MOS was 20.67 and 28.33 for the CG and AR group respectively. Once again, this question has a negative connotation, signifying that the AR group reported that the AR

TABLE IV. QUESTIONNAIRE MAPPING WITH P VALUES AT THE 95% CONFIDENCE LEVEL

QoE Aspect	Question	AR Mean	CG Mean	p^a
Utility	Q1. The instructions were useful.	22.00	27.00	.043
Interaction	Q2. Following the instructions was not interesting.	26.50	22.50	.286
Aesthetics	Q3. I became physically uncomfortable during the experience.	29.42	19.58	.006
Interaction	Q4. My experience was not frustrating.	21.90	27.10	.128
Usability	Q5. I felt confident in my ability to follow the instructions.	23.00	26.00	.355
Efficiency	Q6. Learning to use the instructions correctly was not easy.	24.81	24.19	.859
Usability	Q7. I really enjoyed my experience.	23.77	25.23	.682
Interaction	Q8. The instructions were distracting.	28.33	20.67	.032
Usability	Q9. My experience was stressful.	25.60	23.40	.543
Acceptability	Q10. I would like to experience this form of instruction again.	24.79	24.21	.876
Utility	Q11. Attempting to solve a Rubik's Cube was an enjoyable experience.	23.63	25.38	.577
Usability	Q12. Moving on to the next instruction was easy.	22.38	26.63	.227
Interaction	Q13. Using the instructions felt intuitive.	23.92	25.08	.753
Interaction	Q14. The mode of instruction was not natural.	26.67	22.33	.253

[a] Statistical p-value.

based instructions were more distracting than the paper-based group. This stands to reason because the AR instructions were permanently in the AR user's field of view throughout the experience. Once the CG test subject has read the instruction from the user-manual, they are free to focus their full attention on the task at hand. The trade-off of not having to commit the instruction to short term memory is that the instruction remains in the AR user's field of view, which evidently causes a distraction. This is reflected in the AU results where the AR group had higher classifications of *Annoyance* and in the Q4 results which were next to significant.

Spearman's (r_s) correlations were carried out between the implicit and explicit results. Only the statistically significant correlations are given here. Firstly, for the CG; Q1 correlated negatively to AUs with r_s=-.423, p =.040. Q2 correlated positively to task success rates with r_s=.418, p =.042. Q5 correlated negatively to test duration with r_s=-.466, p=0.22. Q7 correlated negatively to AUs with r_s=-.566, p =.004. Q11 correlated negatively to BPM with r_s=-.495, p =.014. Secondly, for the AR group; Q4 correlated to BPM where r_s=.410, p =.047, and to task success rates with r_s=.457, p=.025. Q5 correlated to AUs with r_s=.436, p =.033. Q6 negatively correlated to BPM with r_s=-.494, p=.014. Q8 also negatively correlated to BPM with r_s=-.522, p=.009. Finally, Q13 and Q14 correlated to SAM arousal levels for the AR group with r_s=.440, p=.031 and r_s=-.509, p=.011 respectively.

VII. CONCLUSIONS

This paper presents a comparative QoE evaluation of AR and paper-based instructions for task assistance. The objective performance metrics show that AR yields efficiency and productivity gains over paper-based instruction. Physiological data showed that the CG users experienced higher increases in skin temperature, heart rate and significantly higher increases in EDA than their AR based counterparts. This suggests that the CG experienced higher levels of arousal than the AR group. SAM questionnaire data shows that the AR users explicitly reported 25% more *Delight* than their paper-based counterparts. The heightened arousal of the CG seen in the physiological data is 25% less '*Delight*' related. A negative correlation of r_s=-.495 between enjoyment (Q11) and BPM corroborates this. This is in line with the findings of [27], where task success rates correlated with the quantity of self-reported '*Delight*'. The questionnaire results showed that the CG group felt that the Rubik's Cube solving instructions were more useful and less distracting than the AR group. A correlation of r_s=.457 between Q4 and success rates shows that the resulting frustration did not impact productivity for the AR group. The AR group reported less comfort than the CG which was due to the AR HMD. Results show that automatically recorded emotional state as expressed through lower facial AUs concur with self-reported emotional state in 21.85% of cases. This demonstrates that the accuracy of non-intrusive emotion estimation using AUs is 4.85% greater than chance. Higher positive valence for the AR group demonstrates that they experienced higher QoE during the Rubik's Cube solving task while achieving higher success rates and lower task completion times. This suggests a link between QoE and productivity. In conclusion, AR HMDs offer evident gains in QoE, efficiency and productivity. However, the current form factor of many AR HMDs may need refined design with user comfort as the motivating factor. Future work will include an analysis of the effects of AR on cognitive load.

ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/16/RC/3918 and the AIT Presidents Seed Fund.

REFERENCES

[1] S. Möller and A. Raake, Quality of Experience, Advanced Concepts, Applications and Methods. Springer, 2013.

[2] L. Hou, X. Wang, L. Bernold, and P. E. D. Love, "Using Animated Augmented Reality to Cognitively Guide Assembly," J. Comput. Civ. Eng., vol. 27, no. 5, pp. 439–451, Aug. 2013.

[3] F. Loch, F. Quint, and I. Brishtel, "Comparing Video and Augmented Reality Assistance in Manual Assembly," in 2016 12th International Conference on Intelligent Environments (IE), 2016, pp. 147–150.

[4] T. Hoβfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in 2011 Third International Workshop on Quality of Multimedia Experience, 2011, pp. 131–136.

[5] C. Keighrey, R. Flynn, S. Murray, S. Brennan, and N. Murray, "Comparing user QoE via physiological and interaction measurements of immersive AR and VR speech language therapy applications.," 25th ACM Int. Conf. Multimed. ACM MM 2017, vol. Thematic Workshop.

[6] D. P. Salgado et al., "A QoE assessment method based on EDA, heart rate and EEG of a virtual reality assistive technology system," in Proceedings of the 9th ACM Multimedia Systems Conference on - MMSys '18, Amsterdam, Netherlands, 2018, pp. 517–520.

[7] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments," in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016, pp. 1–6.

[8] E. D. Demaine, S. Eisenstat, and M. Rudoy, "Solving the Rubik's Cube Optimally is NP-complete," ArXiv170606708 Cs Math, Jun. 2017.

[9] U. Engelke, H. Nguyen, and S. Ketchell, "Quality of augmented reality experience: A correlation analysis," in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017, pp. 1–3.

[10] H. Nguyen, S. Ketchell, U. Engelke, B. Thomas, and P. d Souza, "[POSTER] HoloBee: Augmented Reality Based Bee Drift Analysis," in 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), 2017, pp. 87–92.

[11] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a Research Agenda on Interventions," Decis. Sci., vol. 39, no. 2, pp. 273–315.

[12] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in Advances in Psychology, vol. 52, Elsevier, 1988, pp. 139–183.

[13] S. Sharmila, L. K. Pavithra, and P. Aneerduh, "Rubik's Cube Solution in Augmented Reality Environment," Int. J. Comput. Math. Sceiences IJCMS, vol. 6, no. 8, Aug. 2017.

[14] J. Park and C. Park, "Guidance System Using Augmented Reality for Solving Rubik's Cube," in HCI International 2014 - Posters' Extended Abstracts, Springer, Cham, 2014, pp. 631–635.

[15] T. Rokicki, H. Kociemba, M. Davidson, and J. Dethridge, "The Diameter of the Rubik's Cube Group Is Twenty," SIAM Rev., vol. 56, no. 4, pp. 645–670, Jan. 2014.

[16] I. Wechsung, K.-P. Engelbrecht, C. Kühnel, S. Möller, and B. Weiss, "Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction," J. Multimodal User Interfaces, vol. 6, no. 1, pp. 73–85, Jul. 2012.

[17] G. Paltoglou and M. Thelwall, "Seeing Stars of Valence and Arousal in Blog Posts," IEEE Trans. Affect. Comput., vol. 4, no. 1, pp. 116–123, Jan. 2013.

[18] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," J. Behav. Ther. Exp. Psychiatry, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[19] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler, "Validation of the Empatica E4 wristband," in 2016 IEEE EMBS International Student Conference (ISC), 2016, pp. 1–4.

[20] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A QoE evaluation of immersive augmented and virtual reality speech language assessment applications," in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017, pp. 1–6.

[21] K. Ghamen and A. Caplier, "Positive and Negative Expressions Classification Using the Belief Theory," Int. J. Tomogr. Stat., vol. 17, no. S11, pp. 72–87, Jul. 2011.

[22] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94–101.

[23] "ITU-T P. 913 https://www.itu.int/rec/T-REC-P.913-201603-I/en." [Online]. Available: https://www.itu.int/rec/T-REC-P.913-201603-I/en. [Accessed: 15-Oct-2018].

[24] N. Murray, B. Lee, Y. Qiao, and G. Miro-Muntean, "The influence of human factors on olfaction based mulsemedia quality of experience," in 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), 2016, pp. 1–6.

[25] J. D. Morris, "Observations: SAM: The Self-Assessment Manikin An Efficient Cross-Cultural Measurement Of Emotional Response," p. 6.

[26] M. Valstar et al., "AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge," ArXiv160501600 Cs, May 2016.

[27] K. De Moor, F. Mazza, I. Hupont, M. Ríos Quintero, T. Mäki, and M. Varela, "Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience," presented at the IS&T/SPIE Electronic Imaging, San Francisco, California, USA, 2014, p. 90140U.