Proceedings of the Federated Conference on Computer Science and Information Systems pp. 1631–1640
DOI: 10.15439/2016F535
ACSIS, Vol. 8. ISSN 2300-5963

Limitations of Emotion Recognition in Software User Experience Evaluation Context

Agnieszka Landowska, Jakub Miler
Gdansk University of Technology, Narutowicza St. 11/12, 80-233, Gdansk, Poland
E-mail: {nailie, jakubm}@eti.pg.gda.pl

***Abstract*—This paper concerns how an affective-behavioural-cognitive approach applies to the evaluation of the software user experience. Although it may seem that affect recognition solutions are accurate in determining the user experience, there are several challenges in practice. This paper aims to explore the limitations of the automatic affect recognition applied in the usability context as well as to propose a set of criteria to select input channels for affect recognition. The results are revealed via a semi-experiment based on the case study of an educational game. As a result, a number of concerns were identified, providing a list of pros and cons for affective computing methods applied in the usability testing context. The lessons learned might be interesting for both researchers that develop emotion recognition algorithms and for practitioners, who apply them to diverse areas.**

I. Introduction

ADVANCES in mobile and ubiquitous technologies have made human-system interaction everyday practice in multiple aspects of life. As a result, natural interaction and positive experience of technology is receiving more and more attention. The traditional notion of software usability, as defined by the ISO 9241 standard, includes the effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments [1]. The term *user experience* goes beyond this definition, emphasising the affective component and forming a more holistic picture of human-system interaction [2].

Producers and the marketing/branding industry are interested in the affective aspect of user experience in terms of software products and create a demand for automatic emotion recognition techniques as a tool for getting a larger quantity of more objective data. However, the application of emotion recognition methods in UX testing is not so straightforward. An analysis of only the affective aspect of the user experience might be not enough to determine the issues of effectiveness, efficiency and satisfaction. Therefore Ahn and Picard [3] proposed the affective-behavioural-cognitive (ABC) framework that combines diverse techniques in order to evaluate the user experience in a more holistic manner. The framework was validated with an experiment on beverages [3] and an evaluation of web applications [4]. The ABC framework intends to address goals and interest of diverse stakeholders.

In our research, we develop and study educational games, where user satisfaction is a key affective component, although there are also typical usability issues involved. We turned to the ABC framework as a solution to evaluate the user experience of the software. However, we uncovered many challenges in the practical application of automatic emotion recognition methods. This paper presents the lessons learned and some adjustments in method, that might be useful for researchers who develop emotion recognition algorithms and for practitioners, who apply them in diverse contexts.

The main research questions addressed by this paper might be formulated as follows: *Which emotion recognition and affect representation techniques are applicable within the procedures of usability/user experience testing? What are the main limitations/challenges in their use? How to provide valuable information derived from affective analysis?*

This paper presents a semi-experiment based on a usability case study of an educational game and is organised as follows. Section 2 outlines the previous research on which we based our study. Section 3 includes the operationalisation of the variables and a study plan, while section 4 and 5 provide details of the study execution along with the results. Section 6 provides a summary of the results and a discussion, followed by some concluding remarks (section 7).

Although the authors are aware of the fact, that user experience is a broader term than usability [5], the paper sometimes uses these terms interchangeably, although in the broader (UX) sense.

II. Related work

Work that is mostly related to this research falls into two categories: (1) studies on emotion recognition based on different input channels and their comparison; (2) the use of affect elicitation techniques in user experience evaluation.

(1) There are numerous emotion recognition algorithms, that differ in terms of input information channels, output

This work was supported in part by the Polish-Norwegian Financial Mechanism Small Grant Scheme under the contract no Pol-Nor/209260/108/2015 as well as by DS Funds of the ETI Faculty, Gdansk University of Technology.

labels or representation model and classification method. The most frequently used emotion recognition methods that might be considered when designing an UX evaluation include: facial expression analysis [6], audio (voice) signal analysis in terms of modulation, textual input analysis, physiological signals as well as behavioural pattern analysis [7].

As literature on affective computing tools is very broad and has already been summarised several times, only a sample of papers on recognition methods are provided here. For a more extensive bibliography on affective computing methods, one may refer to Zeng et al. [8] or to Gunes, et al. [9]. The most important conclusions from a review of the literature related to emotion recognition that are the most relevant to this study might be formulated as follows:

(1) Emotion recognition techniques provide results in diverse models of emotion representation (from dimensional models through Ekman's six discrete basic emotions down to two-class classifiers) [10]; there is no common standard model for representing affect;

(2) No input channel is superior to any other in terms of the accuracy and granularity of emotion recognition [11]; a multimodal approach combining diverse input channels provides the most accurate results in most cases; for a multimodal approach, early or late fusion might be considered [12].

(3) Self-report of emotions, although subjective, is frequently used as a "ground truth" (another approach is manual tagging by qualified observers or physiological observations) [13].

The aforementioned results influenced the decisions made concerning the design of this study, especially that it is advisable to use more than one observation channel. The study design is reported in detail in section III.

(2) The second part of the literature review performed under this study was aimed at exploring how automatic affect elicitation techniques are applied in usability and/or user experience evaluation.

There are a few studies on fusing affect recognition and usability evaluation [2][4][12-17]. Most of them consider the usability of food or everyday items and evaluate the overall experience taking emotional factors into account. The most important paper related to this study introduces the Affective-Behavioural-Cognitive approach to UX evaluation [3]. Another is by Lew et al., providing an example of affect evaluation applied to quality assurance procedures for web applications [4].

Kołakowska et al. proposed involving affect recognition in usability evaluation and have suggested four different scenarios: a first impression test, task-based usability test, free interaction test and comparative test [14][15]. The main contribution of the study is the proposal of an emotional state set that might be important in usability evaluation scenarios: *frustration*, *empowerment*, *interest* (*excitement*), *boredom*, *disgust*, *engagement* and *discouragement*.

Partala and Kallinen suggested using the Positive and Negative Affect Schedule (PANAS) scale in self-reporting user experience [16].

Hazlet and Bendek described two studies that used facial electromyography (EMG) measures combined with verbal and performance measures to provide feedback on the user's emotional state. The multimodal approach used in this study was able to provide a measure of the desirability of features, a measure of emotional tension and mental effort expended while performing tasks [17]. There are also some studies of games that utilise the channels used in emotion recognition [18][19][20].

Zimmerman et al. proposed a new method for measuring mood based on the effects of affective processes on motor-behaviour and uses log-files from the mouse and keyboard as a proxy of the mood of the user [21].

There is one study on the limitations on affect recognition in the usability context and it proposed the following criteria: accuracy of emotion recognition, susceptibility to disturbance, independence of human will and interference with usability testing procedures. These criteria were used in an analysis of the recordings from a case study regarding usability evaluation, however were not put into practice [22].

Although some studies on blending affect recognition and usability testing exist, their practical applicability and interference of emotion recognition with IT user experience testing still requires more exploration.

III. Study Design and Research Methods

In order to verify the applicability of emotion recognition in the software UX evaluation context, a semi-experiment was conducted, based on a typical usability study of an educational game extended with user emotion recognition channels. The concept was to use multiple observation channels at the same time, but only those that do not significantly interfere with the typical usability evaluation procedure. Typical usability tests involve 5-10 participants, as this number is enough to reveal 75-90% of usability issues. We planned up to 10 participants for the experiment, as more participants are rarely involved in usability studies.

In order to conduct the study, we chose an educational game, still at the developmental stage, that would be suitable for performing usability evaluation. This choice influenced the participant group. The experiment had to include both the target group of the software under investigation (at least 5 people) and some participants outside the target group since the target group of the application was quite narrow, and we wanted to involve more age groups in the evaluation of emotion recognition techniques.

A. The software under research and UX evaluation goals

GraPM – an educational game about project management [23] was selected for the study. In this game the player assumes the role of a project manager and aims to complete a given product in a given time with some resources under

the uncertainty of some risks. Different product features have different business value, effort and impact on quality. Resources offer different productivity. Threats and opportunities appear and materialize randomly during the development process, requiring the player to take appropriate actions that he chooses from a given set. The effectiveness of particular actions is left for the player to discover during the game-play. Additionally, the satisfaction of the customer and the team must be monitored, as low ratings can result in the abandonment of the project and losing the game. GraPM involves both deterministic and random factors and requires considerable project optimisation to win the game.

The target group of the GraPM game includes two subgroups: (1) students wanting to develop their knowledge and skills in terms of project management; (2) players who enjoy strategy and management games.

The emotional activations that assist in achieving goals were identified as follows: (a) interest – the player should want to learn; (b) slight confusion – the player must be aware that he does not know everything; (c) joy – the player is pleased that he improves and learns; (d) sense of control – the player is content that he can fully control the project/game and win.

The emotional activations that hinder the achievement of goals were identified as follows: (a) fear – the player should not be afraid of learning; (b) strong confusion (frustration) – the player should not be lost and not know what to do; (c) anger – the player should not get angry that he does not understand the game and cannot win; (d) boredom – the game should not be too repeatable and unchallenging; (e) disregard – the player should not consider the game to be of no educational value.

The evaluation of the user experience of the GraPM game is expected to assess to what extent the user experience goals were achieved, with particular focus on learnability. The players should broaden their understanding of the aspects of project management as well as some principles of effective management such as planning, risk management and project supervision. In terms of the game mechanics, the user experience study is expected to provide observations on where the players encounter problems in manipulating the game, which will limit their ability to learn.

The affective extension of the usability study with the emotion recognition should provide additional information on which features of the game enhance learning and which hamper them. Overall, the extended usability study should allow conclusions to be drawn on how to develop the game to improve its educational efficiency, playability, and enjoyment.

B. ABC framework applied in the operationalisation of UX variables

The affective-behavioural-cognitive approach was used in the transformation of the UX study goals into a definition of the semi-experiment variables. We defined the following three general criteria for UX evaluation: understanding, engagement and enjoyment and the criteria were further operationalised into metrics.

Understanding means that the game is comprehensible for a player and this factor corresponds to the cognitive perception of the game mechanics and the game logic (C-cognitive aspect in the ABC approach). According to the information provided on the game, the mechanics understandability should be evaluated after the 2nd and 3rd game, while the understanding of the game logic should be assessed after the 4th and 5th game. Additionally, a learning curve might be derived based on the progress in consecutive game-play.

Engagement indicates that the game is engaging, that it attracts and maintains interest. This factor is a representation of an observable (B - behavioural) aspect of player-game interaction.

Enjoyment determines whether interaction with the game results in a growth in positive affect symptoms, which corresponds to the affective factor (A - affective aspect in the ABC approach).

The author's description of the game provides a list of desirable emotions: interest, slight confusion, joy and feeling of control and a list of undesirable emotional states: fear, strong confusion (frustration), anger, boredom and disregard.

The emotions were listed spontaneously without any guidance or presentation of affect representation models. This approach was chosen purposefully, as a presentation of the models might have influenced the choice. The emotions were mapped into the models provided by the algorithms chosen for the study and the mapping is described in section V.

In this paper we limit our report to the enjoyment factor, although all three aspects were measured and delivered to the game designers as the result of the study [24].

C. Experiment design

In this study we have used the semi-experiment as a research technique. It was a semi-experiment, as it was not possible to fully randomise the choice of subjects to sample. The experiment was based on a real case study, and a group of convenience was used instead of a randomised sample. However, the sample consisted of: representatives of the game target group (students) and some participants outside the group to represent some confounding variables (e.g. age, education and domain). We also set the group size limit (10), as more participants are rarely involved in UX evaluation and this limitation should be taken into account while assessing the affective factor of the game. In other words, one of the challenges was to determine whether 10 people is enough to provide valuable information on affect.

During the study, a limited the number of input channels were recorded (three). Audio (voice) signal analysis and textual input analysis were not considered for inclusion, because in the case study scenario, these channels of human-system interaction were not used. We decided to capture

video image for facial expressions analysis, to ask for self-report based on the PAD emotion representation model and to record physiological signals for reference (skin conductance). The use of other input channels (e.g. keystroke dynamics or mouse usage patterns analysis) are planned in future experiments.

The game-play (which was performed 5 times) was interspersed with questionnaires that measured: competence progress, self-report on emotions, usability questions, including System Usability Scale and questions on the subjective notion of camera and sensor disturbance.

D. Operationalisation of experiment variables

The main goal of the semi-experiment was to answer the research questions as specified in the introduction section — i.e. to determine which emotion recognition techniques are applicable and provide most value in the UX context.

This challenge was conceptualised using a Goal-Question-Metric technique.

GOAL: Analyse the emotion recognition solutions in order to characterise it with respect to applicability from the point view of experimenters relative to the user experience evaluation.

Q1: Is the procedure of software user experience compatible with emotion elicitation techniques?

Q2: Does application of such techniques hinder the process of usability evaluation?

Q3: Does the application of emotion elicitation techniques provide valuable information from the viewpoint of the UX evaluation goals?

These questions are mapped into the following three criteria: applicability, interference and affect-awareness gain.

Applicability represents the degree to which the emotion recognition techniques might be deployed into UX study and the criterion is divided into two factors: input channel availability and susceptibility to noise.

Input channel availability in the UX context was measured by the metric (AP1) time available/time of study ratio.

This study was not focused on the accuracy of classifiers, but rather on the interference (disturbance) introduced by the UX context, as the input channel might be unavailable or significantly noisy.

Susceptibility to noise was evaluated with different proxy metrics for diverse input channels and then qualified to the metric (AP2) level of susceptibility with a common scale of high-medium-low values. The proxy metric for the skin conductance input channel was the number of events that disrupt the channel per time unit (minute) and the events were defined as mouse clicks, which introduced movement artefacts to the EDA signal. The SC sensors (we used two) were placed on the base of the finger and on the wrist [25] and although not all mouse clicks introduced artefacts, most of them did.

For the video channel we used the quality of consecutive frames as the proxy metric.

The ***Interference*** factor measures the influence of emotion recognition application on the usability study. Changes in the usability study (introduced by emotion recognition) should not significantly influence the main goals of the usability study – i.e. gathering information on user effectiveness and learning with software. The factor was measured by 2 metrics: self-report on the subjective notion of camera (IN1) and sensor disturbance (IN2). In the self-report we used a 5-item scale from: 5 - very intrusive to 1 - not intrusive.

Affect-awareness gain is a factor that represents the value of introducing emotion recognition techniques into the software UX context. The criterion is divided in this study into three factors: (AA1) compatibility of the emotion classifier output with emotional states recognition requirements (the ones specified in advance); (AA2) consistency of multimodal observations; (AA3) subjective opinion of the customers of the extended UX study on the value provided by different information on the affective states of the user.

The compatibility metric (AA1) was evaluated for four emotion representation model types (6 basic emotions, arousal only, valence-arousal and PAD positiveness-arousal-dominance models). We used the following scale: 0 – no representation in the chosen model; 1 – could be represented, but might be confused with other emotions; 2 – could be easily and unambiguously mapped; 3 – directly available in the representation model.

Remarks on the consistency of multimodal observations (AA2) were introduced to this study but they will not be evaluated quantitatively. This criterion added, as we observed, huge discrepancies between emotional states estimated on diverse input channels. However, the evaluation of the discrepancy, its scale and analysis of its causes go far beyond the scope of this study. We decided to merely report it, as the differences might compromise the affect-awareness gain.

The results of the players' affect elicitation and analysis were presented to the game designer (the second author of this study) and were evaluated based on the criterion of value they bring to the understanding of the user experience with the game (AA3 metric). The value was measured on a 5-point scale ranging from: (5) very informative to (1) no affect-awareness gain. The designer evaluated the following: the perspectives offered in the presentation of the study results, the views used to visualise data and overall affect-awareness gain in understanding usability and the user experience.

IV. Study execution and the presentation of the UX results

The case study was carried out in 2016 at the Emotion Monitor Stand at Gdansk University of Technology [26]. The The following equipment was used: (1) for physiological signals tracking and analysis: Thought Technology ProComp Infiniti coder, compatible sensors,

Biograph Infiniti Physiology Suite software; (2) for video analysis: a standard Internet camera and video capture software from Logitech, for analysis of facial expressions, Noldus FaceReader was used; (3) for screen capture and user activity tracking and analysis – Morae Recorder, Observer and Manager were used. The three capturing sets were operated at three computer workstations.

We used 2 skin conductance sensors placed on: left-hand fingers and right-hand wrist (for right-handed participants). The locations of the sensors were chosen based on a previous study on the interference of mouse and keyboard usage movements with physiological signals from the fingers [25].

The camera was located above the monitor screen, centrally. Video capture was performed with a 29 FPS rate, 1280x720 resolution and saved as a mp4 file. The analysis of facial expressions was performed using Noldus FaceReader software, providing both Ekman's six basic emotion vectors for each frame as well as valence and arousal model time series.

Morae Recorder was used to capture the screen and gather questionnaire responses and Morae Manager was used to analyse the results.

The study was carried out in April and May 2016. The entire experiment involved 10 participants aged 23 to 43 (8 of them belonged to the game target group), 5 male and 5 female.

The results of the affective aspect of the UX study were reported to the game designer using three perspectives and seven views:

Perspective 1. All UX study participants – information on emotions was summarized for all UX study participants and all tasks performed. The perspective used the following views:

View #1. Declared emotional states versus desired and undesired emotional states

View #2: Recognized emotional states versus desired and undesired emotional states

Perspective 2. Single participant between-task analysis – provides information on fluctuation of emotional states between consecutive gameplays and uses following views:

View #3. Declared emotional states after each gameplay

View #4. Declared/recognized emotional states per gameplay

Perspective 3. Single participant single gameplay analysis – provides detailed information on how emotional state fluctuated during single gameplay, which might be combined with the events. This perspective uses following views:

View #5. Relative frequency of emotional states in Ekman's six basic emotions model;

View #6. Fluctuation of valence (positive/negative state);

View #7. Fluctuation of arousal (calm/active state).

Figure 1 provides sample analytical views on emotion as provided as a result of the UX study.

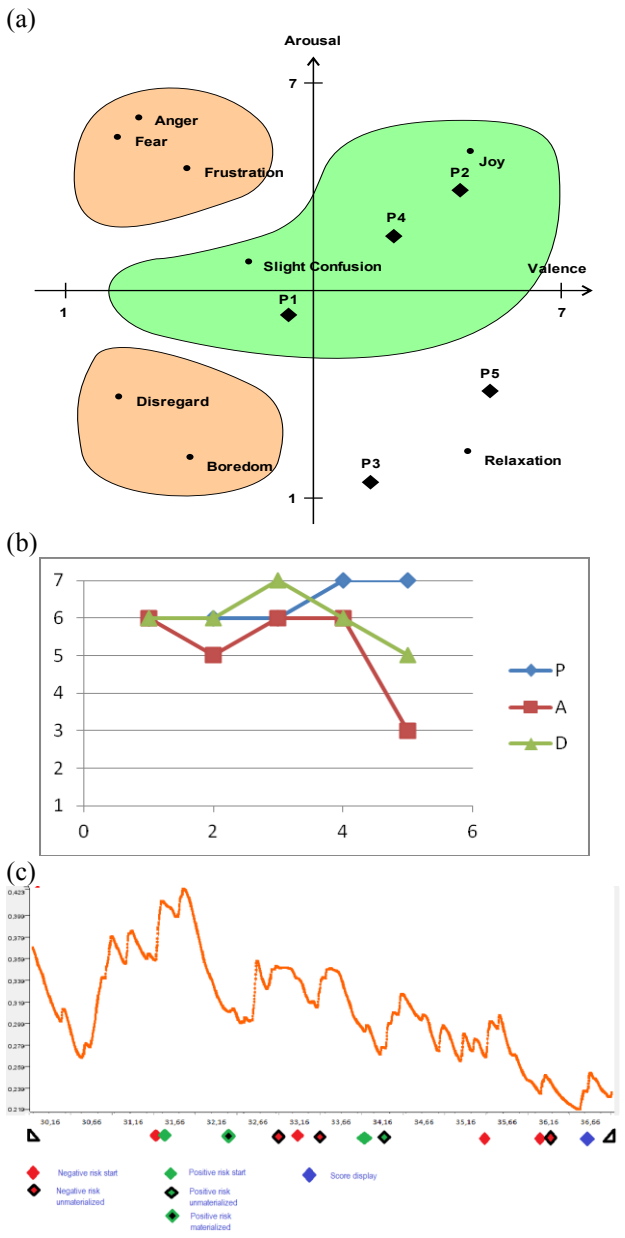

Fig. 1 Sample views from the UX emotion analysis report (a) view #1 (b) view #3 and (c) view #7

View #1 provided in Figure 1 (a) combines various information regarding emotional states. Firstly, it visualises desired and undesired emotional states using the valence-arousal model for emotion representation. The emotions listed by the game designers were clustered into three regions (the scope of the regions was a result of the preliminary mapping of emotional labels to the model and then discussion with the game designer). This preliminary clustering was used in a number of perspectives showing: reported emotional states versus desired/undesired (view #1), recognised emotional states versus desired/undesired (view #2) and one-player emotional state fluctuation from task to task (view #4).

The view provided in Figure 1(b) visualises the change in reported emotional state after consecutive game-plays. It uses the same scale as was used in the questionnaire (1 to 7).

The view provided in Figure 1(c) visualises the changes in the recognised level of arousal accompanied by event markers.

A detailed report on emotional states is not provided here; the views are provided in order to exemplify how the emotions were reported and to show the subject of the evaluation with the affect-awareness gain metrics.

V. Study results

The case study led to some qualitative and quantitative observations. First of all, most of the emotion recognition channels merge naturally with the software usability testing procedure — e.g. the video captured from the camera located near the screen as well as the filling-in of self-report questionnaires on affect. This section reports the results of the study and follows variables as defined in section III.D: applicability, interference and affect-awareness gain. The players participating in the study are coded P1 to P10.

A. Availability and noise susceptibility of the input channels

Applicability was represented with the metric (AP1) time available/time of study ratio and the metric (AP2) level of susceptibility to noise.

The results obtained for the AP1 metric are provided in Table I.

TABLE I.
METRICS OF AVAILABILITY OF THE INPUT CHANNELS

Participant	Self-report (AP1)	Video			Galvanic Skin Response (AP1)
		No of frames	Available frames	AP1	
P1	100%	56313	56313	100%	100%
P2	100%	67392	46323	69%	100%
P3	100%	97354	83601	86%	100%
P4	100%	69325	68512	99%	100%
P5	100%	90101	59275	66%	100%
P6	100%	*	*	*	100%
P7	100%	181135	76423	42%	50%**
P8	100%	56124	54753	98%	100%
P9	100%	70929	70747	99%	100%
P10	100%	69325	65451	94%	100%
Total	100%	na	na	77%	95%

* due to some disk error the video file was corrupted and unrecoverable

** one of the SC sensors detached during the recording session

The self-report on emotional state was merged with the usability study procedure and therefore all participants filled in all 5 questionnaires. Physiological signals were recorded and in one case (P7) only one of the sensors slipped off before the end of the recording. As sensor detachment is a random event, we might assume that such events might occur while hands are used in the human-computer interaction. Video availability varied among participants from 42% up to 100% and this is a result of individual movement patterns. Some participants were seated straight during most of the game-play time. Those with lower video channel availability displayed a number of behaviours that made facial expressions unavailable, e.g. moving sideways (outside the camera range), significant head movement, manipulating the hands in front of the face etc.

The susceptibility to noise (AP2) metrics are provided in table II.

TABLE II.
METRICS OF THE INPUT CHANNELS SUSCEPTIBILITY TO NOISE

Input channel	Proxy metric	Statistics		Susceptibility to noise (AP2)
		Avg	Min	
Video	Frame quality	0,87	0,70	23%
Skin conductance	Time between events causing artifacts [s]	7,79	1,81	80%
Self-report	Subjectivity	na	na	na

Video quality was relatively high and this factor might be improved by using a proper lighting rig as well as improvement in camera resolution. The channel, if available, is quite robust to noise.

One of the prerequisites in recording skin conductance is to restrict the movement of the particular body part that the sensor is attached to. As sensors for skin conductance are placed on the hands, a significant number of movement artefacts occur while using the keyboard and mouse in the UX evaluation procedure. If a keyboard and mouse must be used, the susceptibility to noise is quite high and the condition might be eliminated by moving the sensors to an off-hand location. Feet are mentioned as one of the possible SC locations, although the comfort of the participant might be compromised thereby.

B. Interference with UX procedure

After taking part in the UX evaluation procedure, each participant was asked about the subjective disturbance of the video observation. Camera presence was rated as 1 – non intrusive by 9 out of 10 participants, the other one rated the camera 2 on a 5-point disturbance scale, providing an average disturbance (IN1 metric) equal to 1.1.

We expected that using the camera and screen capture might cause the Hawthorne effect (people behave differently, while being observed), but we did not notice such symptoms.

Physiological measurements require the placement of sensors at the base of the fingers base and on the wrist. Sensor presence was rated as 1 (non -intrusive) by 5 out of 10 participants and 2 (slightly intrusive) by 5 of them Only one rated the sensors 3 on the scale of disturbance so, as a

result, the average disturbance of the sensors (IN2 metric) was equal to 1.6

The result indicates that both sensors and camera were non-invasive, on the whole, for the participants. Some of them even claimed, that they forgot that they were being recorded.

C. Affect-awareness gain

The first aspect of affect-awareness gain is the compatibility of the provided results with the desired/undesired emotions in terms of the affect representation model. The results from facial recognition according to the Noldus FaceReader might be obtained in the form of a 7-item vector of values within <0,1> range corresponding to: anger, joy, disgust, sadness, surprise, fear and neutral state. The results might be also exported to the valence-arousal model of emotions. Physiological signals mainly provide information on arousal and less on valence and therefore should be cautiously interpreted as a labeled emotional state.

The author's description of the game provided a list of desirable emotions: interest, slight confusion, joy and feeling of control and a list of undesirable emotional states: fear, strong confusion (frustration), anger, boredom and disregard. Some of them map directly into the models used by emotion recognition algorithms, and some others require a model of mapping. Table III shows how well the emotions map into diverse models for representation. We used the following scale: 0 – no representation in the chosen model; 1 – could be represented, but might be confused with other emotions; 2 – could be easily and unambiguously mapped; 3 – directly available in the representation model.

TABLE III.
DESIRED/UNDESIRED EMOTIONS AND THEIR MAPPING INTO EMOTION REPRESENTATION MODELS.

Emotion label	Model compatibility (AA1)			
	6 basic +neutral	Valence-Arousal	Arousal only	Valence-Arousal-Dominance
interest	0	1	0	1
slight confusion	2	2	1	2
joy	3	2	1	2
feeling of control	0	0	0	2
Fear	3	1	1	2
strong confusion (frustration)	2	1	1	2
anger	3	1	1	2
boredom	0	2	2	2
Disregard	2	1	1	2
Total	15	11	8	17

Out of the four specified desired emotional states, two are hard to map: interest as a more cognitive than affective state and the feeling of control, which is expressible only with the third dimension of the PAD model – dominance. Some of the desired states are directly available in Ekman's six basic emotions model that includes joy, fear, anger, disgust, surprise and sadness. However, boredom and feeling of control have no representation in this model. Therefore in this study we decided to use the PAD model of emotions. The affect self-report questionnaires were based on this model. For visualisation purpose only, we omitted dominance in some charts (view #1, #2 an #4).

If we use a PAD representation model, the dimensions must be obtained independently from the input channels. The video channel is able to provide valence and some estimation of the arousal, but not in terms of the dominance factor. Physiology-based emotion recognition contributes mainly to the arousal dimension. The only channel that provides the dominance is the self-report. This means that it is difficult to provide one value of emotion estimation reliability. Three independent metrics – one for each dimension should be provided instead.

During the study, we encountered huge discrepancies between the input channels — the self-report and the emotions recognised from the facial expressions were especially contradictory. Sample result showing the inconsistency is provided in figure 2. The figure presents the single-player all-task view #4. Consecutive game-plays were coded as G1 to G5. Diamonds shapes show the reported emotional states, while fuzzy circles depict the recognised ones. The middle of the circle is placed in an area close to the mean recognised state and the fuzziness corresponds to some fluctuations of the recognised emotional state around the average value.

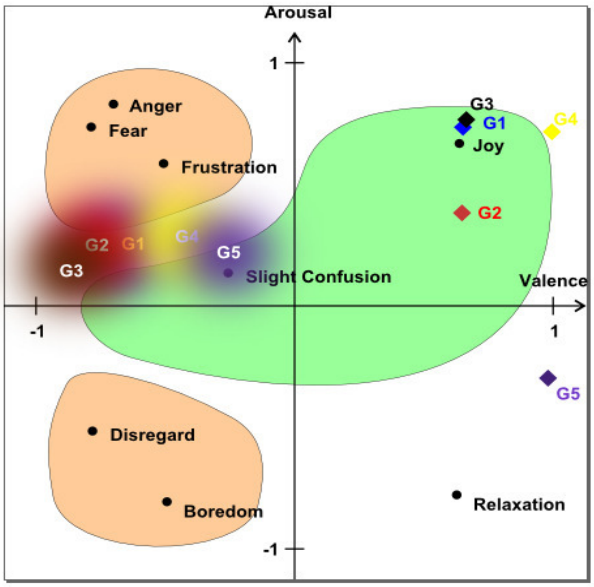

Fig. 2 Single player all-task view illustrating the inconsistency of reported and recognised emotional states

One might observe that the reported emotional states are mainly positive, with varying arousal. The recognised emotional states oscillated in the negative, high arousal quarter (but no extreme arousal levels were observed). The inconsistencies were observed for multiple users and game-plays and the systematic nature of this observation suggests some measurement or recognition error. The consistency of the multimodal observations (AA2) will not be evaluated quantitatively, as this goes far beyond the scope of this study. In trying to explain the reasons for the discrepancy, we watched some sample recordings. We did not notice signs of anger, which was recognised as a dominant emotion in the video. It seemed that perhaps the observable symptoms of concentration (lowered eyebrows) were mistakenly taken as signs of anger. One hypothesis is that the location of the camera above the screen was inappropriate. The inconsistency requires further research and we are planning more experiments to study it.

In this study we adapted to the inconsistency by reporting both recognised and reported emotional states in the UX emotional analysis reports.

The results of the players' affect elicitation and analysis were presented to the game designer (the second author of this study) and were evaluated. The results on the value they bring to the understanding of the user experience with the game (AA3 metric) are provided in table IV.

Some of the perspectives and views provide more information than others. Moreover, it seems that an application of the ABC approach provides more insight into user experience than usability, as was expected.

TABLE IV.
SUBJECTIVE MEASURE OF AFFECT-AWARENESS GAIN

Item type	Item name	Affect-awareness gain (AA3)
Perspective	All UX study participants	4
	Single participant between-task analysis	4
	Single participant single gameplay	2
View	View #1. Reported emotional states versus desired and undesired emotional states	5
	View #2: Recognized emotional states versus desired and undesired emotional states	3
	View #3. Declared emotional states after each gameplay	5
	View #4. Declared/recognized emotional states per gameplay	3
	View #5. Relative frequency of emotional states	2
	View #6. Fluctuation of valence	2
	View #7. Fluctuation of arousal	3
General	Usability understanding	2
	UX understanding	4

Apart from quantitative questions, the survey presented to the game designers included some open questions whereby multiple valuable observations and suggestions were provided:

(1) regarding perspectives: the single-player single-game-play perspective was considered as the least informative, as there were too few in-game events identified to make it interesting; a fourth perspective was proposed to report multiple player single game-play experience;

(2) regarding views: the emotions should be defined using the desired/undesired emotions as listed by the UX goals; some views should have a different scale; a new view should be added that indicates the fluctuation of emotions averaged among the users;

(3) regarding visualisation and reporting: legends and more detailed information should be provided to improve the understanding of the charts;

(4) regarding inconsistencies: the designer tends to believe in self-reporting rather than recognised emotional state, suggesting recognition error.

The most surprising comment was the question: "What is *neutral emotional state*?" This question raises the important issue concerning the proper understanding of the scales and models used for representing emotional states.

One of the expectations, which was not fulfilled in this report, was to provide a chart tagged with events, indicating that certain events caused anger and others caused joy, averaged among the participants. As emotional reactions are very individual, perhaps it would be easier to spot a single nervous system activation than to average the reaction to certain events among the users.

The most important information on affect-awareness gain is that despite the inconsistencies, reporting and visualisation imperfections, the designer claimed that he gained a valuable insight into user experience with the software (rated 4 on a scale from 1 to 5).

VI. SUMMARY OF RESULTS AND DISCUSSION

The main observations revealed through this case study might be summarised with the following statements:

1. It is possible to incorporate emotion elicitation techniques into UX procedures and it seems that emotion recognition has no negative impact on the usability evaluation;
2. Some input channels used in emotion recognition are hard to introduce (e.g. sentiment analysis from text), while others, especially video, self-reporting (and also keystroke dynamics) merge naturally into the UX context;
3. The accuracy of the emotion recognition techniques is compromised by: temporary unavailability of the input channels and susceptibility to noise;
4. Understanding the emotion representation models and mapping the desired and undesired affect to those

models is a preliminary step in providing valuable results.

The practical implications of this study on further applications of emotion elicitation in UX evaluation procedures include:

(1) The advisability to start with the definition of the desired and undesired emotional states and then to map them into one of the representation models for obtaining results.

(2) Selection of the emotion elicitation technique using a set of criteria: availability and susceptibility to noise, possible direct or indirect mapping of the desired/undesired emotional states to the algorithm's output.

(3) As all input channels are subject to temporal unavailability and noise, the challenge might be addressed using a multimodal approach;

(4) While using multiple observation channels, one must look for inconsistencies and the reasons behind them; in the case of discrepancies, perhaps manual tagging by a qualified psychologist might be considered.

(5) Present the results in the form of simple, standard views and provide detailed explanations (assume there is no obvious term regarding emotions).

The results obtained in this study might have some implications for the research on emotion recognition solutions and their integration. The following list of challenges have been identified during this study:

(1) The integration requires a common affect representation model or some mapping between the models. It is quite difficult to integrate and compare results based on labels — a discrete or continuous model might be considered instead.

(2) Emotion recognition algorithms still require improving accuracies and special attention should be given to wild-like conditions, in accordance with current trends in research.

(3) The temporary unavailability of the input channels might be bypassed if the algorithms provide some estimate of the quality of the result (although currently no algorithm does).

We are aware of the fact that the validity of this study has some limitations. We identified and addressed the following threats to its validity: (1) sample size – we engaged 10 users as the usability tests show that 5-10 users reveal 75-90% of the usability issues; (2) sample as a group of convenience – we selected the sample for the UX evaluation to ensure its diversity; (3) confounding variables – we performed the study in a strictly controlled environment, where we limited the possible influences of external factors; (4) subjective measurements – we operationalised most of the variables to objective metrics, the number of subjective self-reports is minimal; (5) observations are based on one case study only – more are planned in the future.

In the future research we would shift to a quantitative approach with results based on a couple of experiments/case studies of UX evaluation procedures based on the ABC framework.

VII. Conclusions

The study revealed three types of challenges: technical, organisational, and related to the cost/value ratio.

Technical challenges (e.g. the accuracy and disturbance robustness of emotion recognition algorithms) might be solved with the future evolution of the affective computing domain, as nowadays emotion recognition in-the-wild conditions is receiving more and more attention. Organisational issues might be eliminated with more experience and trying out different approaches (e.g. multiplying cameras, re-locating sensors). The main challenge remains to provide a reasonable cost/value ratio. Typical usability tests involve 5 to 10 participants, as this number allows 75-90% of usability issues to be revealed. The analysis of emotional states, even when employing automatic affect recognition, is labor-intensive. Multiplying input channels results in higher accuracies, but also introduces the challenge of integration, especially when the observations are contradictory.

Acknowledgment

Authors thank Dominika Makowiecka, who helped in the experiment setting and execution as well as colleagues from the Emotions in HCI Research Group at GUT (emorg.eu), who provided valuable comments on this study.

References

[1] ISO. 1998, Norm 9241: Ergonomics of human-system interaction.

[2] N. Bevan, 2009. What is the difference between the purpose of usability and user experience evaluation methods. In: Proceedings of the Workshop UXEM'09 (INTERACT'09), Uppsala, Sweden.

[3] H.I. Ahn, and R. Picard, 2014. Measuring Affective-Cognitive Experience and Predicting Market Success. IEEE Transactions on Affective Comp. 5(2):173-186, doi: 10.1109/TAFFC.2014.2330614

[4] P. Lew, L. Olsina, P. Becker and L. Zhang, 2012, An integrated strategy to systematically understand and manage quality in use for web applications. Requirements Engineering, 17(4): 299-330 doi: 10.1007/s00766-011-0128-x.

[5] W. Albert and T. Tullis. 2013. Measuring the user experience: collecting, analyzing, and presenting usability metrics. Morgan Kaufmann, USA.

[6] M. Szwoch and P. Pieniążek. 2015. Facial Emotion Recognition Using Depth Data, *The 8th Int. Conf. on Human System Interaction*, pp. 271-277, IEEE, doi: 10.1109/HSI.2015.7170679

[7] A. Kołakowska. 2015, Recognizing emotions on the basis of keystroke dynamics, *Proc. of the 8th International Conference on Human System Interaction*, Poland, doi: 10.1109/HSI.2015.7170682

[8] Z. Zeng, M. Pantic, G. Roisman, and T.S. Huang, 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, , 31(1): 39-58, doi: 10.1109/TPAMI.2008.52.

[9] H.Gunes and B. Schuller, 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions, *Image and Vision Computing*, 31:120-136, doi: 10.1016/j.imavis.2012.06.016

[10] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch and M.R. Wrobel. 2015. Modeling emotions for affect-aware applications. In *Information Systems Development and Applications*, University of Gdańsk, Poland, pp. 55–69

[11] H. Gunes and M. Piccardi. 2005. Affect Recognition from Face and Body: Early Fusion versus Late Fusion, in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, pp. 3437-3443. doi: 10.1109/ICSMC.2005.1571679

[12] I. Hupont, S. Ballano, S. Baldassarri and E. Cerezo. 2011. Scalable multimodal fusion for continuous affect sensing, *IEEE Workshop on Affective Computational Intelligence*, pp.1,8, 11-15, doi: 10.1109/WACI.2011.5953150

[13] J.N. Bailenson, E.D. Pontikakis, I.B. Mauss, J.J. Gross, M.E. Jabon, C.A.C. Hutcherson, C. Nass and O. John. 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses, International Journal of Human-Computer Studies, 66(5): 303-317, doi: doi:10.1016/j.ijhcs.2007.10.011.

[14] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M.R. Wróbel. 2013. Emotion recognition and its application in software engineering, Proc. of 6th International Conference on Human-System Interaction, Poland, pp. 532 - 539, doi: 10.1109/HSI.2013.6577877

[15] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, M.R. Wróbel. 2014. Emotion recognition and its applications, Human-Computer Systems Interaction: Backgrounds and Applications 3. pp. 51-62, Springer. doi: 10.1007/978-3-319-08491-6_5

[16] T. Partala, A. Kallinen. 2012. *Understanding the Most Satisfying and Unsatisfying User Experiences: Emotions, Psychological Needs, and Context. Interacting with Computers,* 24(1):25–34. doi:10.1016/j.intcom.2011.10.001.

[17] R. Hazlett, J. Benedek 2007. *Measuring emotional valence to understand the user's experience of software*, Int. J. Human-Computer Studies, 65:306-314. doi:10.1016/j.ijhcs.2006.11.005

[18] A. Landowska, M.R. Wróbel 2015. Affective reactions to playing digital games, *8th International Conference on Human System Interaction,* IEEE, pp.264-270. doi: 10.1109/HSI.2015.7170678

[19] L. Chittaro, R. Sioni 2014. Affective Computing vs. Affective Placebo: Study of a Biofeedback-Controlled Game for Relaxation Training. International Journal of Human-Computer Studies, 72, 8–9, pp. 663–73. doi:10.1016/j.ijhcs.2014.01.007.

[20] W. Szwoch. 2015. Model of emotions for game players, *8th International Conference on Human System Interactions*, IEEE, pp.285-290. 10.1109/HSI.2015.7170681

[21] P. Zimmermann, P. Gomez, B. Danuser, S. Schar. 2006. Extending usability: putting affect into the user-experience, in *Proc. of Nordic Conf. on Human-Computer Interaction,* Oslo, pp 27-32.

[22] A. Landowska. 2015. Towards Emotion Acquisition in IT Usability Evaluation Context, *Proceedings of the Multimedia, Interaction, Design and Innovation*, 5, doi:10.1145/2814464.2814470

[23] GraPM website, http://grapm.html-5.me.

[24] J. Miler, A. Landowska. 2016. Designing effective educational games - a case study of a project management game, FedCSIS, Gdansk, Poland (accepted)

[25] A. Landowska, 2014. Emotion monitoring - verification of physiological characteristics measurement procedures, *Metrology and Measurement Systems Journal*, Vol XXI, 4:719-732. doi: 10.2478/mms-2014-0049

[26] A. Landowska, 2015. Emotion monitor-concept, construction and lessons learned, in *Proc. of Computer Science and Information Systems (FedCSIS)*, Łódź, Poland, pp.75-80. doi: 10.15439/2015F264