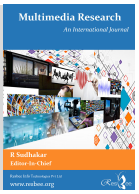
Vol.5 No.3 July 2024

Multimedia Research

Received 26 Mar, Revised 10 June, Accepted 7 July



An Ecological Approach to Measuring User Experience (UX) from Facial Expressions

Zahid Hasan
Shanto-Mariam University of Creative Technology, Bangladesh

Abstract: A system for assessing UX issues automatically is proposed in this paper. The facial behavior of an individual performing a specific activity is tracked in real-time with software that tracks facial motion features. Evaluated with the conventional studies, this approach has several advantages: ease of deployment in the user's natural setting; avoidance of invasive devices; and severe cost minimization. An evaluation of the user experience of the system was conducted using 144 videos that showed 12 users executing three tasks on four commercial media players. To predict the presence/absence of UX issues based on the tracker's features, we used different machine learning algorithms. We show promising outcomes that open up opportunities for automated real-time UX estimation in an environmental context.

Keywords: User; Automated; Facial; Media; Tasks

1. Introduction

In recent years, HCI has taken a more holistic approach to user experience, including emotional variables alongside standard usability requirements. Consequently, new instruments are needed to measure emotional responses to interactive technologies. In contrast to emotion research using physiological measures, user-experience (UX) evaluation settings require ecological validity, and self-report methods adapted from psychology or developed ad-hoc for addressing emotional responses to interaction technologies don't have significant validation.HCI presents new challenges to emotion researchers, as it tends to elicit low-intensity emotional reactions that are not often accompanied by visually observable changes [2]. Emotional reactions are also typically mixed and cannot be captured by simple emotion taxonomies such as those implemented in emotional tracking tools like the FaceReader by Noldus Vision [6]. A central feature of HCI is that emotions change over time, while the majority of psychological and marketing research has focused on reactions to static stimuli that do not interact with their perceivers [5].HCI is further characterized by dynamic emotions that change over time, Finally, researchers in HCI are interested in emotions as a way to understand the quality of interactions. In order to collect physiological measurements, they may not possess the theoretical and methodological background they need. An automatic tool for detecting the emotional status of users during technology interaction, eMUST (eMotional Unit System), is presented in this paper as a preliminary validation. As it requires neither invasive devices nor constant lighting settings, our system fulfills the ecological requirement of UX research and thus can be used anywhere with a commercially available video camera. Additionally, eMUST allows users to distinguish between mixed emotions, such as frustration, by tracking minute changes in facial muscle activity. Additionally, the system requires no special knowledge of emotions and is cheap. The study reports and discusses the results of extensive experiments carried out on 12 users executing three tasks on four commercial media players.

2. Related Work

We considered different methods of judging a user's emotional state when interacting with technology. Several tools were used to generate self-reports. PAD indicates pleasure arousal and dominance emotional scale [7], an Independent three-dimensional measurement of emotions using semantics such as pleasure, excitement, and dominance differential scale. The self-assessment scale (SAM) is the Evolution of the PAD image used to measure Emotional reactions to various technological products

https://doi.org/10.46253/j.mr.v5i3.a4

Services such as interfaces and interactive television content [8] and the current user interface [9]. It turned out to be SAM is more efficient than the original counterpart in measuring Individual responses to emotional stimuli [10]. PreEmo [3] is Software that allows users to evaluate the emotions they evoke in product design using a series of cartoon characters displayed on the screen. Premo has 14 emotions (7 positive, 7 negative) in product design and they can be combined to describe complex emotions. Another well-established one Tools - Geneva the Wheel of Emotions (GEW) has been developed [12] based on a group of 20 categories of emotions, available in several languages. Overall, self-reports share the advantage to be ecologically valid, as they do not require to be administered in a controlled setting. On the other hand, the information provided by these instruments can be interpreted in different ways in different cultures, although graphical and animated versions (e.g., SAM, PrEmo) can alleviate the problem to some extent.Rating scales are also subject to 'response bias [13], by which participants tend to respond to surveys according to the answer they think the experimenter wants, rather than according to their true feelings or opinions. Another drawback is the dissipative nature of emotions; since users report their emotional experience after the interaction with the system they may not remember how their emotions developed over time. Finally, even small differences in wording can raise the cognitive noise level and modify response patterns. An alternative approach to self-report deals with real-time measurement of motor expression and physiological arousal of emotional response. Psycho-physiological arousals, such as changes in blood volume pressure, heartbeat rate, brain activity, skin conductance, and motor expressions in charge of eye movement, facial expressions, or vocal tone can be measured through various apparatus (i.e. electrodes, sensors, diode). Yet, the face is a favorite target. Studies that evaluated the emotional component of users' interaction with technology through the observation of patterns of facial movements relied mainly on two techniques: the detection of facial muscle activity through electromyography (EMG), and the extraction of facial features through software for video signal processing. Branco et al. [14] recorded the electrical activity of three muscle groups (corrugator, frontalis, and zygomatic) while participants performed a formatting task. A significant correlation was found between task difficulty and the muscle activity of the corrugator (eye-brow down-wards) and the zygomatic (mouth). EMG signals (related to zygomatic and corrugator muscles) were also used in [15], alongside verbal and performance measures. The outcomes exhibited that EMG values were effective in providing real-time affective measures and that muscular activity was generally congruent with verbal data. Facial EMG was exhibited to be an effectual model to track emotional changes over time, but this technique provided information only on emotional valence and not on specific emotions. Ravaja et al. [16, 17] studied the psychophysiological response to different digital game events through EMG signals of the corrugator, zygomatic and orbicularis muscles. Results showed that stimuli generally considered positive and stimuli perceived as negative were associated with positive emotional responses. At the same time, passive reception of negative stimuli was associated with low-arousal negative emotions. [18] examined the reliability of 'FaceReader', a software adopted by Vicar Vision and Nold us Information Technology on the basis of Ekman and Friesen's theory of the Facial Action CodingSystem (FACS) distinguishing six fundamental emotions, on usability evaluation [21]. They collected data about the user's emotional state from three sources: self-reporting questionnaires, researcher's logging, and data from FaceReader. The outcomes exhibited consistency from the data provided by FaceReader and expert-human judgments when questionnaire data were not consistent with the other sources of emotional information. Current research in HCI follows the assumption that no single method is enough to assess emotions effectively. [19] conducted an experiment to assess web page design with a multi-method approach. They employed four methods: self-reports facial electromyography (EMG), galvanic skin response, and eye-tracking. They found that EMGof corrugator muscle was more active on some specific types of web pages, such as 'Facts' pages than 'About' pages.They argued that this was because 'Facts' pages contain some 'hard-hitting' text. The experiment described in [20] confirmed that systems with maximum usability tend to have more positive emotions than systems with usability flaws. Different types of emotion were measured using a variety of methods: SAM, electro-dermal activity, heart rate, EMG of corrugator and zygomatic muscles, the time required for input operations, and the Geneva appraisal questionnaire. The authors found a small correlation among the components of the emotion triad (subjective feeling: SAM, motor expression: EMG, physiological reaction: Heart rate). The facial expression did not correlate with physiological measures, but both facial expressions and physiological measures correlated with SAM. Results also showed that the activity of the zygomaticus major was not a consistent indicator of positive feelings. Overall, psycho-physiological measures can capture users' emotions in real-time and they are language-independent; however, they are costly, require a controlled environment, and are intrusive.

TABLE 1: Relation between Action Units, Motion Units, and activated muscles.

MU	AU(s)	Description	Activated Muscle(s)
1	10	Upper Lip Raiser	Levator labii superioris
2	16	Lower Lip Depressor	Depressor labii inferioris
3	20	Left Mouth Corner Horizontal Deformationt	Risorius
4	12, 13	Left Mouth Corner Vertical Deformation	Zygomatic major Levator anguli oris
5	20	Right Mouth Corner Horizontal Deformation	Risorius
6	12, 13	Right Mouth Corner Vertical Deformation	Zygomatic major Levator anguli oris
7	1, 2, 4	Right Eyebrow Deformation	Frontalis; Corrugator Depressor glabellae Depressor supercilii
8	1, 2, 4	Left Eyebrow Deformation	Frontalis; Corrugator; Depressor glabellae Depressor supercilii
9	6	Right Cheek Deformation	Orbicularis oculi
10	6	Left Cheek Deformation	Orbicularis oculi
11	5, 7	Right Lid Deformation	Levator palpebrae superioris Orbicularis oculi (pars palpebralis)
12	5, 7	Left Lid Deformation	Levator palpebrae superioris Orbicularis oculi (pars palpebralis)

In the following frames, a template-matching model is exploited to evaluate the two-dimensional motion of the interest mesh nodes. By projecting the two-dimensional motion information onto the three-dimensional face model, the mesh is updated. eMUST brings advantages to EMG and other approaches to video analysis as it does not need persistent devices and can be exploited in natural settings, involving the situation with decisive or varying illumination circumstances; at the same time, it uses fine-grained facial features tracking in preference to relying on a fixed emotion classifier. This feature permits for more precise detection of low-intensity, mixed emotions as the ones elicited in HCI. Our optimized implementation is able to run in real-time (¿25 frames per second), and requires no specialized hardware; an 800 MHz processor, along with 512 Mb of RAM and an OpenGL-capable graphic card is sufficient for real-time operation. In a different context, [25] exploited facial motion features in order to detect personal highlights in movies. To the best of our knowledge, this paper reports initial exploratory work combining an ecological experimental setup along with fine-grained tracking of facial features in the usability evaluation context.

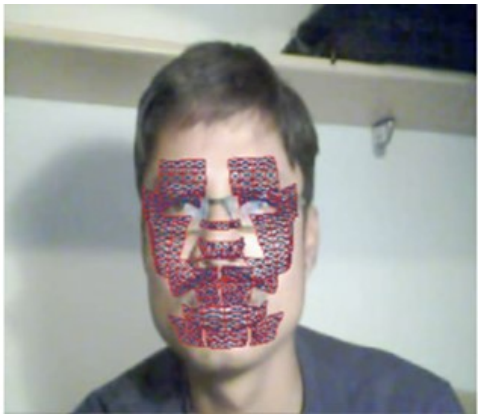


Fig.1. *Snapshot of eMUST in action*

3. Method

To investigate the potentialities of eMUST as a tool for user experience evaluation, we conducted an experiment in which participants' faces were recorded by the webcam of their laptops while they were using and evaluating different media players. The video data served as input for eMUST. The information on users' facial expressivity was then compared with questionnaire data and with judgments

on participants' facial expressivity as an indicator of task difficulty provided by human observers. Furthermore, the video data were used to train our system to predict user experience problems based on facial emotional prompts.

3.1 Participants

In the experiment, 15 Masters' Students (14 M, 1 F; mean age = 26,4 years) from a local University were included, voluntarily, as participants. All the participants declared to have a minimum of three years of experience in exploiting media players. No particular inclusion criteria were applied to the sample. Before accepting to participate in the experiments, participants were informed that the study involved audio and video recordings of their faces and voice and that they had the right to remove from the experiment at any time.

3.2 Settings and Materials

The experiment was conducted in an ecological setting, which could be the participant's room at the Student Dormitory or one of the offices at the university. In the study, the materials exploited were four freeware media players ("iTunes, Music Bee, Media Monkey, and Songbird") (Fig. 2), a folder containing a list of songs, a slideshow presentation containing one screenshot for each one of the interfaces of the four media players, and four identical questionnaires, one for each media player, for the assessment of the user's experience. The alternative option of the four media players was on the basis of the fact that none of the participants had preceding experience with any of them. The four media players were installed on the participants' laptops or PCs, besides a program for audio and video recording (BB Flashback Express) that allowed recording the signals acquired by means of the inbuilt webcam and microphone. The same software was also used to record all users' actions on the interface. Before the study, the experimenter checked that all the software was correctly installed and working.

Fig.2. *The Four media players: iTunes(a), Music Bee(b), Songbird (c), and Media Monkey(d)*

4. Procedure

Before the experiment, participants explained the general intention of the research and were asked to sign an approval form. Initially, the analysis initiated with the analyzer exhibiting to participants the screenshots of the four media players shown in arbitrary order for 10 seconds, to gather an initial-impression assessment of the interface. At the time of the next step of the study, participants were questioned to carry out similar three tasks on the four media players and to fill in the questionnaire of UX instantly subsequent to the conclusion of the tasks on one media player. The order in that participants exploited the four media players were randomized when the order of the three tasks was similar for all the participants. The three tasks are a: importing a folder comprising songs to the media library of the player; b: examining a specific song in the media library and playing it; c: adjusting the song equalization by exploiting the equalizer inbuilt in each media player. Once participants finished the

fourth questionnaire, they were questioned to select one of the four media players and to entrust to utilizing it for the subsequent month rather than their normal program. During this time period, participants were surveyed at regular intervals asking them to fill in the UX questionnaire. During the study, the evaluator remained in the room with the participant to interfere in case of technical issues, however, the number of interactions with the participant was reserved for the least amount.

5. Description of the Media Players

To perform the tasks on each of the four players, different sequences of actions were necessary. There was a difference in the look and feel, usability features, and functionality between the four media players used in the study. The tasks differ in their level of difficulty and in the least amount of steps needed to finish them (Table 2). In task 1 For Media Monkey and Music Bee, the process of importing a media folder was less intuitive than Songbird and iTunes. For instance, Songbird, as well as iTunes, featured the words “add folder” and “Import” in the menu, and Music Bee and Media Monkey exploited the word “Add/Rescan” which may be confusing for the user. Also, Songbird and iTunes aided drag-and-drop to import media folders. For task 2, the actions sequence needed (find and play a song) does not present significant differences among the media players. In all of them, songs can be sorted along with title, artist, album, genre, and date; and also all players featured an internal search engine. Nevertheless, they differed regarding the visual illustration of the songs, with iTunes featuring the most appealing interface. Task 3 (applying and locating equalizer) was simple in Songbird, iTunes, and Media Monkey than in Music Bee. This difference was because, in the first three media players, the equalizer was positioned under one menu (“view, control, or tool”) from where it must be directly evaluated. Music Bee contains an equalizer under the “view and controls” menus. Choosing it from the controls menu does not generate any visible reaction on the system. If you wish to view the Equalizer menu, you must select both the 'player control panel' and then 'show equalizer' items. At this point, the equalizer window opens and needs clicking a radio button to enable it. Both Music Bee and Media Monkey have an icon on the most important interface to access the equalizer quickly, but these icons are complex to comprehend and observe.

5.1 Pre-processing of video data

The video clips recorded during the experiment were manually examined and annotated and all errors concerning task performance were marked. Errors were operationallydefined as a number of individual variations from the optimal performance, i.e., the procedure requiring the minimum number of actions following the strategy adopted by the user. Undesired user movements due to conversation with the experimenter and the user’s spontaneous gestures that obstructed their faces were also marked. Then the video was exported into xvid format and a separate text file containing that annotation information was generated in order to label the MU feature vectors extracted by eMUST.

5.2 Dependent variables

The experiment included four classes of dependent variables: a) Performance measures related to task execution; b) judgments regarding user experience presented by participants via questionnaires; c) human observers’ judgments about the difficulty of the tasks evinced by video analysis, looking at facial expressions of the participants; d) facial cues extracted from the videos using eMUST.

6. Performance Metrics

Errors were computed by subtracting the number of steps carried out by the participant and the minimum number of steps required to finish the task. Seconds were recorded to determine the completion time of the task.

7. User Experience Evaluation

Participants were asked to answer three parts of the questionnaire: their assessment of media players, information regarding their preceding use of media players, and fundamental demographic information. The evaluation of media players included individual dimensions of UX (aesthetics, usability, pleasure, symbolism, and functionality), as well as summary judgment. Four items were used to measure usability: Simple to use, simple navigation, suitable use, and simple orientation [26]. A two-factor technique of aesthetics was exploited that differentiates between classical and expressive aesthetics [26]. Traditional

aesthetics indicates conventional notions of beauty emphasizing symmetry, order, and clear design. It involves 5 items: Symmetric design, Clear design, clean design, Pleasant, and Aesthetic design. Expressive aesthetics is characterized by qualities that capture the user's perception of the creativity and design originality: original, sophisticated design, creative design, use of particular effects, and fascinating Symbolismconcerns the inference of connotative meanings related to an interactive device [27]. As opposed to aesthetics, symbolism assessment is based on cognitive processes. It involved the following items: fits personality generates positive associations, indicates likable things, communicates desirable images, and presents a positive message regarding the user. The pleasure was evaluated by 3 items presented in [26]: feel pleasure, feel joyful, and feel gratified. Four components were evaluated for functionality: completing the tasks required, producing expected results, interfacing with other software, and providing security. Summary judgments were based on 5 items from [28] and [29]: I will employ it in the future, I would suggest it to a friend, It will be fun to use, I feel I will be required to have it, and I will be fulfilled with it Because of the small sample size, it was not possible to validate the scales psychometrically. For each UX dimension, dependent variables were calculated by averaging appropriate items, as enlighted by a large corpus of preceding research [26,29].

8. Human Observers' Judgments

Three independent observers (2 M, 1 F; mean age: 27 years) were included to understand how well people can detect usability issues by seeing the faces of users carrying out the tasks. The set of 144 video clips (12 participants * 4 media players * 3 tasks) gathered at the time of the experiment was divided into six blocks of 24 clips each. The clip order was randomized in a way that each block comprised clips recorded by all the participants. All three observers gazed at and rated all six blocks of videos; the presentation order of the blocks was randomized among users. For each video clip, observers were questioned to rate "how difficult is the task the person shown in the video is dealing with"; judgments were expressed by exploiting a 5-point Likert scale in which the value 1 corresponded to "very easy" and the value 5 corresponding to "very difficult".

9. Facial cues Extraction Process

The original videos collected during the study were processed. De-noising was achieved during data pre-processing by using median filtering in a sliding window approach. Median filtering is a widely used de-noising technique consisting of 1) defining a window size; 2) centering the window on each element of the vector to be de-noised; 3)replacing the current element value with the median value of the elements in the window. For each subject, mean and standard deviation values of all motion units were computed. Then, we counted how many times the absolute values of a motion unit went above one standard deviation over the mean of that motion unit for the duration of 1/16 seconds (frame duration). This computational model was based on the process recommended in [15], which justifies the use of one standard deviation interval based on the standard in psychological testing, wherein these values are considered to be out of the normal range. Following [15], a measure of elevated tension was associated with the values of MU7 (right eyebrow deformation) and MU8 (left eyebrow deformation). Measures of frustration and confusion were based on the work of [30], who found a correlation between the AU12(lip corner puller) and frustration, and the combination of AU4, AU7, and AU12 and confusion. Frustration was calculated by calculating the vector length resulting from the adding of the respective vertical vectors (MU4/MU6) and horizontal (MU3/MU5), according to the Pythagorean theorem. Confusion was calculated using the quadratic mean of the individual features as captured by eMUST (eyebrows, eyelids, and mouth corners).

10. Results

A 168 tasks sample was gathered and involved in the subsequent analyses, wherein the task was exploited as the unit of analysis. Per task, the average number of errors was 3.98 (stddev = 7.14), ranging from a "0" minimal to a "39" maximal. Some 48% of the tasks contained no errors, 26% of them contained less than 5 errors, and the remaining 26%contained 6 or more errors. The distribution shape of errors was enhanced by calculating the square root of each data point. This normalized variable was evaluated by an ANOVA with media-player (4) and task (3) as the between-subjects factor. Post- hoc comparisons were on the basis of the Least Significance Difference approach. The most important effect of media player ($F(3,156) = 6.73$, $p < .001$, partial $\eta 2 = .12$) and task were significant ($F(2,156) = 7.83$, $p < .01$, partial $\eta 2 = .09$). Descriptive statistics of the original error scores are reported in Table 3. The post-hoc

analysis recommended that the media player effect was because of the poor performance of Musicbee while no differences occurred among the other media players. The effect of task was because of task 2 which was significantly simpler than the other two tasks. The outcomes of the Anova on the square root of time highlighted a very similar pattern. The main effects of media player ($F(3,156) = 5.85$, $p = .001$, partial $\eta 2 = .10$) and task were significant ($F(2,156) = 7.70$, $p < .001$, partial $\eta 2 = .09$). Post-hoc analysis highlighted that Musicbee took the longest time to perform the tasks (with no significant differences between the other media players), and task 2 was significantlyfaster than the other tasks. A high linear correlation between error rate and execution time was found ($r(168) = .88$, $p < .001$).

10.1 Judgments about UX

Scores to the 6 UX dimensions tested in the study were entered as dependent variables into a set of repeated-measures analysis of variances, with a media player (4) and time (2) as within-subjects factors. The factor time referred to the evaluation phase: questionnaires were collected after a 10 seconds exposure to the interface and after actual interaction with the system. The outcomes of the analyses are summarized in Table 4. The analyses returned a very coherent framework of results, featuring an important most important effect of the media player and time for all variables, with no interaction. The effect of time was due to a constant decrease in the evaluation scores after actual usage. The media player effect reflected different evaluations of the systems. In all UX dimensions, I-tunes and Media-monkey were perceived and evaluated as the best systems, with no significant differences between them. In terms of usability, functionality, pleasure, and overall judgment, Songbird and Musicbee scored more negatively than each other. These trends of results are only partially consistent with the participants' ultimate selection of the media player to employ for the subsequent month. Indeed, 10 people were determined to exploit i-tunes, whereas Media Monkey (despite high scorings on all UXdimensions) was chosen only by 2 people. The remaining3preferred Songbird.

TABLE 2: Means and standard deviations (in bracket) for performance data and usability evaluation.

Media Player	Task1 Error	Task1 Time	Task2 Error	Task2 Time	Task3 Error	Task3 Time	Usability Score
iTunes	2.71 (3.65)	30.3 (25.4)	1.00 (1.88)	30.9 (28.3)	4.14 (7.97)	24.0 (27.5)	5.45 (1.12)
Songbird	4.36 (6.51)	33.4 (29.3)	0.93 (1.69)	15.7 (11.1)	3.21 (3.75)	13.6 (11.4)	4.90 (1.17)
MediaMonkey	5.50 (8.39)	37.5 (32.4)	0.21 (0.43)	21.5 (15.0)	1.93 (3.22)	14.1 (13.5)	4.50 (1.33)
Musicbee	10.36 (9.70)	69.6 (61.8)	2.57 (4.84)	27.9 (39.2)	10.79 (13.3)	53.2 (51.1)	3.57 (1.29)

10.2 Correlation between MUs, usability errors, and subjective evaluation

The correlations between the individual motion features in the literature have been identified as being related to task difficulty or negative emotions and the number of errors evinced in the task is reported in Table x. There is a strong correlation between the different features and errors, particularly for eyebrows movements (corresponding to the corrugator muscle activity), lids, and cheek movements(corresponding to orbicularis oculiuscle). The correlation between the mouth and the errors is relatively lower. The composite variables frustration and confusion were alsosignificantly correlated with errors, in the order $r = .38$ $p < .001$ and $r = .50$ $p < .001$. Subjective estimations gathered by questionnaires at the conclusion of the third task exhibited little correlation with error numbers. "The values for functionality ($r = .20$) pleasure ($r = -.15$) and overall judgment ($r = -17$) were important. Fascinatingly, all these correlations were because of the participant's performance at task 3. Task 1 and 2 returned no correlations at all, wherein task 3 reported higher and important correlations for all variables however expressive aesthetics". No correlations were identified between questionnaire data and emotional variables.

11. Prediction Experiments

We experimented with three machine-learning approaches in order to evaluate the predictive power of the automatically extracted facial motion features on the occurrence of errors. The machine learning algorithms used to train the different models were Convolutional Neural Networks(CNN), K-Nearest Neighbor Classifier (KNN) Support Vector Machine (SVM) and Multi-Layer Perceptron (MPL), and naive Bayes classifier (NB). The learned models were tested in two experimental scenarios. The first scenario is a user-dependent environment, where models were trained and evaluated for each individual user. The second scenario used a user-independent model. In the latter case, given the whole set of features, a generic model was trained and evaluated independently of the users. In the user-dependent scenario, each session (i. e. one user performing three tasks on four media players) was divided into four datasets;

each dataset contained the MU features extracted for a given media player. For each user, four models have trained on the features extracted from three media players and then tested against the remaining dataset. The prediction results were then averaged. In the user-independent scenario, the whole dataset was taken into consideration and divided by 12 (user number); for each user, a model was trained on the features extracted from all the other users and tested against the features extracted during his session. These scenarios allowed us to evaluate our system in a realistic context of use. In one case it would be trained on a given user and evaluated on new observations regarding a new piece of software. In the other case, it would be trained on a set of users and evaluated on new observations regarding an unseen user. As an evaluation measure, the F-Measure (F1score) is used. The F1 score is defined as the harmonic mean between precision (number of correct outcomes divided by the size of the test set) and recall (number of correct results divided by the number of outcomes that must have been returned). It ranges from 0 (all examples wrongly classified)to 1 (perfect match). In our case, since we wanted to evaluate the system in an online scenario, an approach similar to cross-validation was adopted: instead of randomizing the whole dataset and dividing it into almost equal parts, we chose to divide the data on a media-player basis in one scenario, and on a user basis in the other. This ensures that a given MU feature vector extracted from a media player/user is either entirely used for training the model or for testing it, thus simulating a production scenario where a model would be trained on previous tasks performed by the same user or on the previous users. In the learning stage, we evaluated the models trained with the MUs extracted by the automatic system and labeled them according to the ground-truth annotation. Table 5 and Table 6 show the results of these dependent and user-independent experiments using the entire set of MU features at our disposal. As a threshold to separate positive and negative examples, we used values of 1 (any task completed with one or more errors is marked as negative) and values of 5 (any task completed with fiveor more errors is marked as negative). By analyzing the results in the user-dependent scenario we can infer that users have different thresholds for visually displaying frustration. When users make only one error, they may display little or no visual sign of frustration. When making more errors users might be gaining awareness of their difficulty to accomplish the task and be more prone to visually displaying frustration. Furthermore, there are users (e.g., user 11) who are not inclined to show frustration and might be unfit for emotional research. In the user-independent scenario, we tested how well the learned models would perform against an unseen participant. Results are not satisfactory, as could be expected by the variance of visual reaction of the participants even when presented with the same task. As common sense suggests, people react with different dynamics, when presented with the same situations.

TABLE 3: Summary of Questionnaire Ratings.

Dependent Variable	Source	df	F	P	Partial
Usability	Media Player	3.14	11.02	.001	.44
	Time	1.14	9.69	.01	.41
	MP*Time	3.14	0.88	ns	–
Classical aesthetics	Media Player	3.14	12.75	.001	.48
	Time	1.14	8.19	.05	.36
	MP*Time	3.14	0.38	ns	–
Expressive aesthetics	Media Player	3.14	.38	.001	.39
	Time	1.14	8.79	.01	.43
	MP*Time	3.14	10.44	ns	–
Functionality	Media Player	3.14	7.55	.001	.35
	Time	1.14	8.59	.05	.38
	MP*Time	3.14	.68	ns	–
Symbolism	Media Player	3.14	11.20	.001	.44
	Time	1.14	7.82	.05	.36
	MP*Time	3.14	2.21	ns	–
Pleasure	Media Player	3.14	11.56	.001	.45
	Time	1.14	5.10	.05	.27
	MP*Time	3.14	2.21	ns	–
Summary judgement	Media Player	3.14	9.86	.001	.41
	Time	1.14	8.75	.05	.39
	MP*Time	3.14	3.14	ns	–

TABLE 4: Correlation values between MU features and UX errors.

dsd	R– mouth corner vertical	L– mouth corner vertical	Right Eyebrow	Left Eyebrow	Right Cheek	Left Cheek	Right Lid	Right Lid
errors	.444 **	.376 **	.673 **	.742 **	.498 **	.632 **	.628 **	.467 **

TABLE 5: User-dependent predictions' F1 scores; models trained using all MU features.

User	MLP	KNN	NB	SVM	CNN
2	55.4	62.4	61.5	70.95	80.1
3	59	71.9	54.45	76.66	83.5
4	34.6	44.75	67.03	71.33	81.1
5	73.65	100	57.13	100	100
6	54.25	78.15	60	50.2	60.6
8	73.13	34.2	61.43	30.83	40.7
9	55.65	53.55	56.9	61.58	80.8
10	65.68	33.35	56.8	35.74	50.7
11	35.27	23.34	46.93	46.6	60.7
12	50.15	58.1	69.85	73.65	90.6
13	74.55	64.35	79.38	81.1	90.5
14	82.38	54.3	47.25	75.98	85.3
average	59.48	56.53	59.89	64.55	84.1

TABLE 6: User-independent predictions' F1 scores. Models trained using all MU features.

MLP	KNN	NB	SVM	CNN
58.1	50.74	51.92	61.83	80.2

12. Human validation

Out of the three judges who were asked to score the videos, only one performed well against the ground truth. "The scoring was highly correlated with the number of errors ($r = .48$, $p < .001$) and with the facial features as extracted byeMUST". This person was the only one who has had specifictraining on FACS.

13. Conclusions

This paper reports evidence of correlations between emotional facial features and people's performance while using three commercial media players. In particular, movements of the eyebrows were found to be powerful indicators of task difficulty, whereas mouth expressions were weaker. During times of tension, participants tend to touch the lower part of their face, causing video occlusion and hampering the automatic detection of their emotional states. The prediction experiments performed on the user-dependent scenario show encouraging results: error occurrence could be predicted by looking at the individual emotional display to a variable, yet a reasonable level of probability. This feature could be exploited in UX studies by training the algorithms with a set of extremely short tasks (in the order of 10 seconds) to gain knowledge about their facial behavior, and automatically detect errors during actual testing. The training tasks should be carefully designed to involve different levels of complexity and elicit different forms of emotions. This phase could also be used as a screening for participants, as our data demonstrated that emotional reactions are subjected to very high inter-individual variability. The Convolutional NeuralNetwork (CNN) method was the one that performed better in real-life scenarios. Yet its performance can be greatly improved by taking a few strategies, such as adopting a more complex weighting method and ensuring reliable and clear examples for training. These strategies will allow reducing the number of outliers in the training set, thus leading tosignificantly improved performance. Prediction results in the user-independent scenario do not show good results because of the high degree of between-subjects variance concerning the visual display of frustration that machine learning algorithms can hardly deal with. This finding is consistent with [25], where recorded facial motions of several users presented with a series of short movie clips showed very different trends. Despite its intrinsic limitations using a triangulation of different techniques, we believe that automatic facial expression analysis can be an extremely valuable tool to support user experience studies. In particular, it can help to monitor the dynamic evolution of emotions during the time, and counteract the tendency evinced in the present study whereby the questionnaire reflected the performance of the very last tasks, rather than providing a summary judgment of the overall experience., an automatic tool becomes a necessary help to untrained evaluators when they need it. The findings of our study show that it is

extremely difficult to infer performance behavior from visual cues and requires an expert level of specialization. Indeed, our results show this dissipation performance Behavior based on visual cues is extremely difficult and needs a high level of specialization. Unlike [18] we maintain that UX research requires methods that can be applied beyond the usability laboratory in this regard, our approach represents a unique selling point and promising tool.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Hassenzahl, M., & Tractinsky, N. (2006). User experience-a researchagenda. Behaviour & information technology, 25(2), 91-97.

[2] Benedek, J. and Miner, T.: Measuring Desirability: New methodsfor evaluating desirability in a usability lab setting. In: Proceedingsof Usability Professionals Association, pp. 8-12 (2002)

[3] Desmet, P.M.A.: Measuring emotion; development and applicationof an instrument to measure emotional responses to products.In: M.A. Blythe, A.F. Monk, K. Overbeeke, & P.C. Wright (Eds.),Funology: from usability to enjoyment. pp. 111-123. Dordrecht:Kluwer Academic Publishers (2003)

[4] Katherine Isbister, Kia H¨o¨ok, JarmoLaaksolahti, and MichaelSharp.: The sensual evaluation instrument: Developing a transcultural self-report measure of affect. Int. J. Hum.-Comput. Stud.65(4), 315-328 (2007)

[5] Derbaix, C.M.: The impact of affective reactions on attitudes towardthe advertisement and the brand: A step toward ecological validity.Journal of Marketing Research. 32(4), 470-479 (1995)

[6] Den Uyl, MJ and Van Kuilenburg, H.: The FaceReader: Online facialexpression recognition. In: Proc. Measuring Behaviour, pp. 589-590.(2005).

[7] Mehrabian, A.: Pleasure-arousal-dominance: A general frameworkfor describing and measuring individual differences in temperament. Current Psychology. 14(4), 261-292 (1996)

[8] Chorianopoulos, K. and Spinellis, D.: User interface evaluation ofinteractive TV: a media studies perspective. Universal Access in theInformation Society. 5(2), 209-218 (2006)

[9] Swindells, C. and MacLean, K.E. and Booth, K.S. and Meitner, M.:A case-study of affect measurement tools for physical user interfacedesign. In: Proceedings of Graphics Interface 2006, pp. 243-250.Canadian Information Processing Society (2006)

[10] Bradley, M.M. and Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry. 25(1), 49-59, Elsevier(1994).

[11] Millard, N. and Britton, R.: Calling time: an effective and affectiveevaluation of two versions of the MIT Beer Game. In: Proceedingsof the 21st British HCI Group Annual Conference on HCI 2008:People and Computers XXI: HCI... but not as we know it-Volume2, pp. 75-77. British Computer Society (2007)

[12] Scherer, K. R.: What are emotions? And how can they be measured? Social Science Information, 44(4), 693-727 (2005)

[13] Dawes, R.M.: Fundamentals of Attitude Measurement, Wiley, NewYork, NY (1972)

[14] Branco, P., Firth, P., Encarnao, L.M., Bonato, P.: Faces of emotionin human-computer interaction. In: CHI'05 extended abstracts onHuman factors in computing systems, pp. 1236-1239. ACM (2005)

[15] Hazlett, R.L. and Benedek, J.: Measuring emotional valence tounderstand the user's experience of software. International Journalof Human-Computer Studies. 65(4), 306-314 (2007)

[16] Ravaja, N. and Saari, T. and Laarni, J. and Kallinen, K. and Salminen, M. and Holopainen, J. and Jarvinen, A.: The Psychophysiologyof Video Gaming: Phasic Emotional Responses to Game Events.DIGRA Conf. (2005)

[17] Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., Keltikangas, J.:arvinen, L.: The psychophysiology of James Bond: Phasic emotionalresponses to violent video game events. Emotion 8(1), 114–120(2008)

[18] Zaman, B. and Shrimpton-Smith, T.: The FaceReader: Measuringinstant fun of use. In: Proceedings of the 4th Nordic Conference onHuman-computer interaction: changing roles, pp. 457-460. ACM(2006)

[19] Westerman, S., Sutherland, E., Robinson, L., Powell, H., Tuck, G.: AMulti- method Approach to the Assessment of Web Page Designs.In: Proceedings of the 2nd international conference on AffectiveComputing and Intelligent Interaction, pp. 302-313. Springer-Verlag(2007)

[20] Mahlke, S. and Minge, M., Thuring, M.: Measuring multiplecomponents of emotions in interactive contexts. In: CHI'06 extendedabstracts on Human factors in computing systems, pp. 1061-1066.ACM(2006)

[21] Ekman, P., & Friesen, W.: Facial Action Coding System: A tech-nique for the measurement of facial movement. Palo Alto, CA:Consulting Psychologists Press (1978)
[22] Ryan, A., Cohn, J. F., Lucey, S., Saragih, J., Lucey, P., la Torre, F. D.,and Ross, A.: Automated facial expression recognition system. In:IEEE International Carnahan Conference on Security Technology(2009)
[23] Kuilenburg, H. V., Wiering, M., and Uyl, M. D.: A model basedmethod for automatic facial expression recognition. In: Proceedingsof the European Conference on Machine Learning, pp. 194–205(2005)
[24] Tao, H., and Huang, T.: A piecewise bezier volume deformationmodel and its applications in facial motion capture. Advances inImage Processing and Understanding. 39–56 (2002)
[25] Joho, H., Staiano, J., Sebe, N., and Jose, J.: Looking at the viewer:analysing facial activity to detect personal highlights of multimediacontents. Multimedia Tools and Applications. 51, 505–523 (2011)
[26] Lavie, T. and Tractinsky, N.: Assessing Dimensions of PerceivedVisual Aesthetics of Web Sites. International Journal of Human-Computer Studies. 60(3), pp. 269-298 (2004)
[27] Tractinsky, N. and Zmiri, D.: Exploring attributes of skins aspotential antecedents of emotion in HCI. Aesthetic computing. 405-422 (2006)
[28] Lund, A.: Measuring usability with the USE questionnaire. Usability and User Experience. 8(2)(2001)
[29] De Angeli, A. Sutcliffe, A. Hartmann, J. (2006).: Interaction, usability and aesthetics: what influences users' preferences? In: DISConference Proceedings, pp. 271-280. ACM (2006)
[30] McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K.Graesser, A.: Facial Features for Affective State Detection in Learning Environments. In: CogSci07, pp. 467-472, (2007)