UX_Mate: From Facial Expressions to UX Evaluation

Jacopo Staiano, María Menéndez, Alberto Battocchi, Antonella De Angeli, Nicu Sebe
Dept. of Information Engineering and Computer Science, University of Trento
Via Sommarive, 5 – 38123 Povo (TN), Italy
{staiano,menendez,battocchi,deangeli,sebe}@disi.unitn.it

ABSTRACT
In this paper we propose and evaluate UX_Mate, a non-invasive system for the automatic assessment of User eXperience (UX). In addition, we contribute a novel database of annotated and synchronized videos of interactive behavior and facial expressions. UX_Mate is a modular system which tracks facial expressions of users, interprets them based on pre-set rules, and generates predictions about the occurrence of a target emotional state, which can be linked to interaction events. The system simplifies UX evaluation providing an indication of event occurrence. UX_Mate has several advantages compared to other state of the art systems: easy deployment in the user's natural environment, avoidance of invasive devices, and extreme cost reduction. The paper reports a pilot and a validation study on a total of 46 users, where UX_Mate was used for identifying interaction difficulties. The studies show encouraging results that open possibilities for automatic real-time UX evaluation in ecological environments.

Author Keywords
User eXperience, evaluation, video analysis, mixed emotions, facial expressions.

ACM Classification Keywords
H.5.2. [**Information interfaces and presentation**]: User interfaces – e*valuation/methodology*.

General Terms
Experimentation; Human Factors; Performance; Design.

INTRODUCTION
In the last decade there has been a widespread move in HCI to consider emotional aspects of User eXperiences (UX) alongside the standard usability requirements [17]. This move has brought forward a need for new instruments to measure emotional responses to technology. Psychologists have long striven to overcome the difficulties of operationalising and measuring emotions, yet the HCI context introduces new complex challenges. Self-report instruments [2,12,19] are in need of a serious validation effort, and invasive physiological instruments contrast with the requirement of ecological validity of the evaluation settings. The measurement is further complicated by the low intensity emotional reactions often elicited in HCI settings [2,12]. These reactions tend to be of a mixed nature [12] and are normally not accompanied by visually observable changes in a person state [11]. As such, they are difficult to be described using the basic emotion taxonomy [14] implemented in current tools for usability evaluation [10]. Furthermore, HCI researchers and practitioners are interested in emotions as a means to understanding dynamic interactions, whereas the bulk of research in psychology and marketing has considered static stimuli [12]. Finally, most HCI practitioners are likely to miss the theoretical and methodological background necessary to interpret self-reports or to operate complex and expensive physiological instruments.

This paper presents UX_Mate (UX Motion Activation Tracking Engine), a tool for the automatic detection of the dynamic user emotional state during interaction with technology. The system fulfils many requirements of UX research: (i) it does not need invasive devices nor controlled illumination settings; (ii) it can be installed in any device featuring a commercial in-built video camera; (iii) it tracks minute changes in the facial muscle activity of the user facilitating discrimination of mixed emotions, such as frustration or confusion, and (iv) it is cheap and does not assume heavy background knowledge. The paper presents two independent evaluation studies used to validate the performance of UX_Mate against that of skilled user researchers. The paper focuses on usability evaluation of different interactive devices through facial cues; the approach can be extended to cover mixed feelings such as frustration, a feeling linked to interaction difficulties as in our scenario, flow and fun. The main contributions of this research are: (*a*) the development and evaluation of UX_Mate; and (*b*) a corpus[1] of synchronized and annotated videos of interactive behaviour and facial expressions, which can be used to ground research on the relationship between behaviour and emotion in HCI.

RELATED WORKS
A large corpus of research has explored the computational implications of technology that 'relates to, arises from, or deliberately influences emotions' [28]. However, less

[1] The corpus is available at disi.unitn.it/~staiano/trentoux

emphasis has been devoted so far to understanding how measures of emotions can support the evaluation of interactive devices, and to the validation of new measurement tools. In this section, we summarise the state of the art of different approaches to emotion appraisal in UX evaluation.

Questionnaire measures

Dozens of affective inventories are available in the psychological literature. Questionnaires share the benefit of being ecologically valid, as they do not need to be administered in controlled settings. However, they can only provide a summary evaluation of past events and cannot capture the dynamics of the interaction. Due to the dissipative nature of emotion, this evaluation is likely to be affected by response bias. One of the most extensively used questionnaires is PAD [26], which measures emotions on three independent dimensions (Pleasure, Arousal, Dominance) by means of a semantic differential scale. Although PAD is reliable, there are a number of difficulties associated with it. Firstly, it requires the respondents to provide 18 different ratings for each stimulus. Secondly, it requires statistical skills from the evaluators. Finally, the cultural frame of the respondent can bias verbal ratings: even small differences in wording can increase the level of cognitive noise and alter response patterns.

To alleviate these problems, HCI research has recently focused on shorter, non-verbal measurement tools. The ones most commonly used in evaluations [3,6,33] rely on visual representations of emotions. Examples are the Self-Assessment Manikin [3] and PrEmo [12]. Some research has also investigated the communication of emotion through tactile experiences with physical stimuli [19]. Yet, questionnaires may still have several issues, as many dimensions of user experience are not stable, singular judgments, but rather vary over the time course of the interaction.

Psycho-physiological measures

A number of psycho-physiological measures, such as changes in blood volume pressure, skin conductance, heart beat rate, brain activity, and muscular activity responsible for changes in facial expressions, eye movement, or vocal tones, can be measured through various devices (e.g. sensors, electrodes, diodes). Facial expressions are a rich source of information through which people convey emotion. Research in psychology demonstrated that facial expressions show reliable correlation with self-reported emotions [8] and with physiological measures of emotion [9]. The most common approach used to measure patterns of facial movements in HCI relies on the detection of muscle activity through electromyography (EMG) [4,18,29,36,24]. Such studies investigate the electrical activity of several muscles (corrugator, frontalis, orbicularis and zygomatic) in a range of interaction tasks. Results are preliminary and at times contradictory, but overall they suggest a relationship between the activity of the corrugator (eyebrow movement) and zygomatic (mouth corner movement) with interaction events. Overall, facial EMG was showed to be an effective method for tracking emotional changes over time. Yet EMG is not the expected panacea to the measurement requirements of HCI as it tends to provide exclusively information on emotional valence and does not provide clear information on the specific emotions elicited. Furthermore, there are still issues of external validity: facial expressions and self-reports do not always correlate [24,37].

While physiological approaches share the benefit of being able to accurately capture changes in emotional states that cannot be measured using other methods [31], they all require specific expertise as well as special and expensive equipment [23]. To overcome these limitations, researchers started investigating how usability can be assessed by means of automatic analysis of facial expressions collected by video signal processing. A commercial system is FaceReader, developed by VicarVision and Noldus Information Technology [10]. This tool, based on Ekman and Friesen's theory [14] of the Facial Action Coding System (FACS), can recognize six basic emotions (i.e., happiness, sadness, anger, disgust, surprise, and fear), which are returned as output of the video analysis. The system was tested in a usability evaluation triangulating data from three sources: questionnaires, human judgments, and data from FaceReader [37]. The results showed consistency between FaceReader's output and expert-human judgment, while questionnaire data were not consistent with the other sources of emotional information. This lack of correlation can be due to the direct use of basic emotions, which are unlikely to be elicited in the HCI context. Moreover, FaceReader has a number of constraints related to illumination or background clutter, which can affect the output [15]. Although the first results obtained using video analysis are encouraging, further research is needed to face current limitations, with a particular concern about finding ways of exploiting psycho-physiological measurements in a cheap, non-invasive, and ecological fashion.

UX_MATE

UX_Mate (UX Motion Activation Tracking Engine), is a software tool developed for automatic assessment of UX by means of facial motion tracking. UX_Mate brings together the advantages of EMG and approaches based on video analysis since it does not require invasive devices and can be used in natural settings, including situations with critical or varying illumination conditions. Moreover, it exploits fine-grained facial motion tracking instead of relying on a fixed emotion classifier. This feature allows to take advantage of low-intensity, mixed emotions as the ones elicited in HCI. UX_Mate does not focus on emotion recognition: it rather exploits global and local facial motion patterns building on a framework based on the anatomical analysis of the human face derived from FACS [14].

Since it was first proposed in 1978, FACS has been established as the most widely accepted coding system for facial expressions in a number of different research contexts and has been widely employed in vision based automatic analysis of human faces [10,21,27,30]. FACS provides definitions for over 40 Action Units (AU), that correspond to the contraction or relaxation of one or more muscles of the face and are responsible for facial appearance changes. The subtle motion of facial muscles corresponding to fast transitory motion of AUs is a powerful indicator of micro-expressions [13], i.e., the involuntary expressions appearing for periods of time as short as 1/25 of a second. A distinctive property of such micro-expressions is that they can hardly be faked [13].

Despite this large success, FACS presents some limitations: human observers require specific training in order to exploit it [13] and it is very time consuming: coding 1 hour of video data requires 4 hours of work [5]. UX_Mate overcomes this limitation by a tracking system able to run in real-time. As opposed to other approaches, the system is robust to illumination changes.

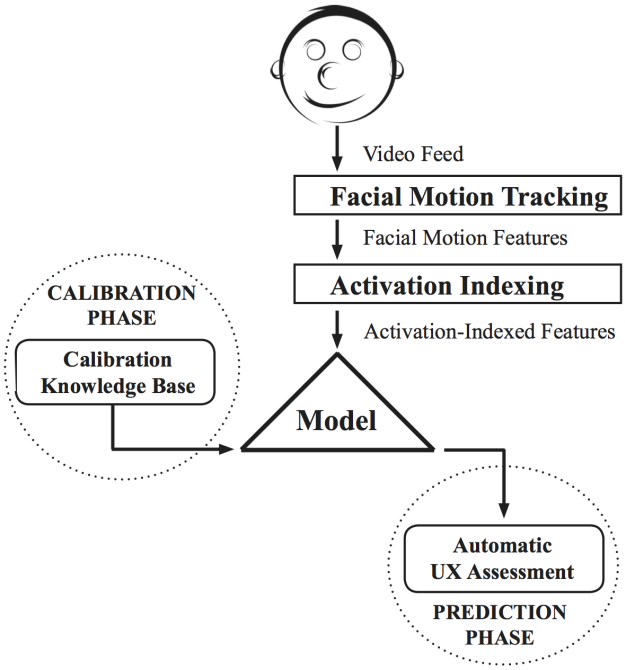

Figure 1. System overview of UX_Mate

A graphical overview of UX_Mate is portrayed in Figure 1. The video serves as input to the facial motion tracking module, whose output is then post-processed by the Activation Indexing module. The data generated by the Activation Indexing module serves as input to a generic Machine Learning module: in the calibration phase the activation-indexed features are used, in conjunction with the knowledge of the calibration task they refer to, to build a model of the user's reactions. Such model is then used, in the prediction phase, to automatically assess UX in the tasks under evaluation. Manually labeled data is used in this paper to evaluate UX_Mate's performance.

Facial Motion Tracking System

The facial motion tracking system endorses a framework inspired by FACS: 12 Motion Units (MU) are defined in correspondence to one or more Action Units. The tracking information refers to the movement of these 12 MUs, corresponding to a subset of AUs defined in Table 1. This subset has proven to be sufficient for automatic facial expression recognition [7]. The key difference between AUs and MUs is that the latter not only represent activation of the unit(s), but also magnitude and direction of motion, making the measurement more informative.

MU(s)	**Facial Movement** ***Activated Muscles***	AU(s)
1	Upper Lip Raise *Levator labii superioris*	10
2	Lower Lip Raise *Depressor labii inferioris*	16
3/5	Right/Left Mouth Corner horizontal Deformation *Risorius*	20
4/6	Right/Left Mouth Corner vertical Deformation *Zygomatic major, Levator anguli oris*	12/13
7/8	Right/Left Eyebrow Deformation *Frontalis, Corrugator, Depressor glabellae, Depressor supercilii*	1/2/4
9/10	Right/Left Cheek Deformation *Orbicularis oculi (pars orbitalis)*	6
11/12	Right/Left Lid Deformation *Levator palebrae superio glabellaeris, Orbicularis oculi (pars palpebralis)*	5/7

Table 1. Description of Motion Units

The tracking implementation we employ in UX_Mate is an adapted version of the algorithm proposed in [32]. This algorithm makes use of a Piecewise Bezier Volume Deformation (PVBD) model in a fixed camera environment; video data is captured by webcam. In the initialization stage, a near-frontal and neutral face is captured and a 3D facial mesh model is fitted on the face. Such a model consists of sixteen surface patches (which are guaranteed to be continuous and smooth) embedded in Bezier volumes. The control points of the surface patches are the facial points of interest represented by the MUs.

On the chosen initialization frame, the motion vectors are set to zero; on subsequent frames, a template matching method is used to estimate the two-dimensional motion of the mesh nodes of interest. The mesh is then updated by projecting the two-dimensional motion information onto the three-dimensional face model. For each processed video frame, the facial motion tracking module outputs a 12-dimensional vector. The values in the output vector correspond to the direction and magnitude of motion for the corresponding MUs. We will refer to such values with the term “features” from now on.

For the sake of the current studies, we computed two variables based on the combination of different MUs, namely, confusion and frustration. The algorithm was based on FACS based research [25] showing a correlation between AU12 (lip corner puller) and frustration, and between the combination of AU4, AU7 and AU12 and confusion. Frustration features were computed for both mouth corners by calculating the length of the vector resulting from the addition of the respective horizontal (MU3/MU5) and vertical vectors (MU4/MU6). Confusion was computed by the quadratic mean of the individual MUs (eyebrows, eyelids, and mouth corners). Additionally, a measure of the overall facial activity was computed by adding all motion units values. New variables can be added to cover a larger set of feelings linked to interaction events of interest. The described tracking algorithm has been previously used in several application scenarios, e.g. for affective video summarization [38].

Activation Indexing

The purpose of the Activation Indexing module is to detect MUs' activation at each frame, and subsequently score each task in terms of activation levels. The module computes the mean and standard deviation for each motion unit. The Activation Index for a MU is given by the count of frames (1/25 of a second) where its absolute value goes above one standard deviation over the mean. Such a computational approach is based on the procedure suggested in [18]. The criterion of one standard deviation is justified by standard in psychological testing, where these values are considered to be out of the normal range.

Machine learning module

The activation-indexed features are used as input data for a generic machine-learning module written in the Java programming language and based on the WEKA [16] data mining tool. The modular approach adopted in designing UX_Mate makes it possible to connect data output to any machine learning toolkit with a Java interface. This module returns a prediction on the level of occurrence of a given event.

A high level of flexibility is achieved through the use of *calibration tasks*, i.e. short sessions carefully designed in order to elicit specific reactions according to the goal of the evaluation to carry out. In the example described in this paper, the participants were tested with two short tasks designed to elicit a variable level of difficulty, but the tasks can easily be updated to fit different requirements.

PILOT STUDY

A pilot of UX_Mate was run during the evaluation of four Media-Players. The videos of participants' faces served as input for UX_Mate, which returned feature values representing the level of activation of the motion units and the compound variables tracked. These values were used as predictors of performance (interaction time and errors). Furthermore, a measure of external validity was collected asking three human observers' to judge the difficulty of the tasks based on the videos of the users' faces.

Method

Participants

Fifteen Masters students (14 M; mean age = 26.4 years) of a local University were involved, on a voluntary basis, as participants in the study. All of them reported at least three years of experience with different media players, but none had ever used the ones tested in the evaluation.

Procedure

The study was conducted in several places, including rooms in the hall of residence or university offices using the participants' own computers. Four Media-Players were tested: iTunes, MusicBee, Songbird, and MediaMonkey. They each had a different look-and-feel, and level of usability and functionality. The media-players were installed on the participants' laptops alongside a program for synchronized video and audio recording of faces and screen actions.

Before the study, the participants signed a consent form stating that their face would be videotaped but with no reference to why. Then they performed three tasks on each media player: importing a folder to the library; finding a song and playing it; adjusting the equalization of a song. Media player order was counterbalanced across participants, while task order was kept constant. After task completion, participants filled in the UX questionnaire referring to the media player they had just used. At the end of the study, they chose one media player and committed to use it instead of their usual program for the following month. During the experiment, the evaluator remained in the room with the participant in order to intervene in case of technical problems, but the amount of the interaction with the participant was kept to a minimum.

Measures

The study collected three classes of measures: performance data, self-reports on user experience, and facial cues extracted by UX_Mate. Performance data (errors and time) were obtained from the expert analysis of the interaction videos. Users' interaction path for each task was compared to an ideal task analysis describing optimal performance (i.e., the procedure allowing reaching the goal with the least number of actions). Several analyses were performed for each task and media players to account for different possibilities to achieve the same goal (e.g., direct manipulation vs. menu selection). All user actions not matching optimal performance were counted as errors.

The questionnaire was composed of three parts addressing UX evaluation, information about participants' previous usage of Media-Players, and demographic data. Media players were evaluated for individual dimensions of UX

and summary judgement. A definition of each dimensions and item wording is reported in Table 2. Dependent variables were computed averaging items of individual scales (all $\alpha > .80$).

Based on literature analysis, we selected 4 MUs (MU4/6 describing the movement of the zygomatic major and MU7/8 describing the movement of the corrugator) which could better describe the facial expressions of people facing difficulties. Furthermore, we analysed the compounded indexes describing frustration and confusion.

Usability	interaction quality [22] *Easy to use, easy navigation, convenient to use, easy orientation.*
Classical Aesthetics	traditional notions of beauty emphasising symmetry, order and clear design [22] *Clear design, Symmetric design, Clean design, Pleasant, Aesthetic design.*
Expressive Aesthetics	perception of creativity and originality of the design [22] *Original, sophisticated design, creative design, use of special effects, fascinating.*
Symbolism	inference of connotative meanings associated to an interactive device [34]. *Fits personality, creates positive associations, represents likable things, communicates desirable image, provides a positive*
Pleasure	enjoyment in the interaction [22] *Feel joyful, feel pleasure, feel gratified*
Functionality	evaluation of system capability *Performs the tasks required, produces expected results, can interact with other software, can provide security*
Summary judgment	behavioural intention *I will use it in the future, I would recommend it to friend, It will be fun to use, I feel I will need to have it, I will be satisfied with it.*

Table 2. UX scales.

Results

Performance data

A sample of 168 tasks was collected and used for performance evaluation. The average number of error per task was 3.98 (SD= 7.14), ranging from a minimum of 0 to a maximum of 39. The distribution shape of the variable error and time was improved by computing the square root of each data-point. The normalized variables were analyzed by two ANOVAs with media-player (4) and task (3) as between-subjects factor. Post-hoc comparisons were based on the Least-Significance Difference method. Partial eta-squared (η_p^2) was used as an estimate of effect size.

The ANOVA on error returned a significant effect of media-player ($F_{(3,156)} = 6.73$, $p < .001$, $\eta_p^2 = .12$) and task ($F_{(2,156)} = 7.83$, $p < .01$, $\eta_p^2 = .09$). The ANOVA on time returned a significant effect of media-player ($F_{(3,156)} = 5.85$, $p < .001$, $\eta_p^2 = .10$) and task ($F_{(2,156)} = 7.70$, $p < .001$, $\eta_p^2 = .09$). In both analyses, post-hoc tests indicated that MusicBee was significantly worse than all other systems (with no significant differences between them), and that task 2 was significantly easier than the other tasks.

Questionnaire data

The scores of the 6 UX dimensions tested in the study were entered as dependent variables in 6 repeated-measures analysis of variances, with media-player (4) as within-subjects factor. The analyses returned a significant main effect of media-player for all variables. The F values ranged from 7.55 ($p = .001$, $\eta_p^2 = .35$) for the functionality score to 12.75 ($p < .001$, $\eta_p^2 = .48$) for classical aesthetic.

iTunes and Media-monkey were preferred in all UX dimensions, with no significant differences between them. Songbird and MusicBee scored negatively, with significant differences favoring SongBird on usability, functionality, pleasure and summary judgement. This trend of results is consistent with participants' final choice of the media-player to use for the next month. Ten people decided to use iTunes, nobody selected MusicBee.

Performance/questionnaire correlation

A correlation analysis was performed on total number of error and UX dimension scores. The correlation matrix reported only 2 small but significant negative correlations for functionality and classical aesthetics ($p < .05$). The analysis was repeated for each individual task, showing that the number of errors committed in task 1 and task 2 were not correlated with any UX dimension. The number of errors in task 3 showed 4 significant correlations, for usability, classical aesthetics, functionality and pleasure.

Facial cues

Three separate ANOVAs were performed to analyse the conjoint effect of the MUs measuring activity in the right or left hand-side of the face. In particular, we analysed MU4 and MU6 (eyebrow deformation), MU7 and MU8 (mouth corner movement), and the two frustration index as a function of media-layer (4) and task (2). The analysis on the MUs related to the zygomatic muscle (MU4 and MU6) returned a main effect of media-player ($F_{(6,254)} = 2.22$, $p < .05$, $\eta_p^2 = .05$). Post-hoc analysis indicated significantly more activation of both MUs during the interaction with MusicBee. The analysis on the MUs related to the corrugator (MU7 and MU8) returned no significant effects.

The analysis on frustration returned a multivariate main effect of media-player ($F_{(6,258)} = 3.04$, $p < .01$, $\eta_p^2 = .07$). The difference was due to MusicBee, who elicited more frustration than all other systems. Identical results emerged from the Anova on confusion. The effect of media-player was significant ($F_{(6,258)} = 5.26$, $p < .01$, $\eta_p^2 = .11$) and MusicBee was identified as the worse system.

Correlations between the facial features and the number of

Facial Movement (MU)	Num. Errors
Right mouth corner vertical Deformation (MU4)	.444**
Left mouth corner vertical Deformation (MU6)	.376**
Right Eyebrow Deformation (MU7)	.673**
Left Eyebrow Deformation (MU8)	.742**
Frustration right	.49**
Frustration left	.28**
Confusion	.50**

Table 3. Correlation between MU and errors. ** p < .01

errors are reported in Table 3. There was a strong correlation between eyebrows movements (corresponding to the corrugator muscle activity) and the number of errors. The correlation between the mouth and the errors was lower, but still significant. A regression analysis was run to investigate the influence of the 4 motion units in predicting the number of errors. Using the enter method a significant model emerged ($F_{(4,140)}$ = 63.2 p < .0010). The model explained 65% of the variance. Table 4 gives information for the predictor variables entered in the model. MU8 was the strongest contributor to the model, MU4 and MU6 were also significant but contributed substantially less in the prediction.

MU	B	SE B	ß
MU4	.004	.001	.17**
MU6	.004	.001	.13*
MU7	.004	.003	.12
MU8	.013	.002	.61**

Table 4. Regression coefficients * p < .05; ** p < .01

Human Validation

Three independent observers (2 M, 1 F; mean age: 27 years) were involved in an external validation of UX_Mate. The objective was to understand the level of agreement between UX_Mate and the human capability to detect task difficulty by observing the faces of users performing the tasks. The videos collected during the experiment were divided into six blocks of 24 clips each. The order of clips was randomized in a way that each block contained clips recorded by all the participants. The three observers watched and rated all the six blocks of videos; the order of presentation of the blocks was randomized. For each video clip, the observers were asked to rate “how difficult is the task the person shown in the video is dealing with”; judgement were expressed using a 5-point Likert scale. Only one judge performed well against the ground truth (i.e., actual numbers of errors in the interaction). Her scoring was highly correlated with the number of errors (r =.48, p < .001) and with the facial features as extracted by UX_Mate (r ranging from .19 to .33). This person was the only one who has had specific training on FACS. The other evaluators' scorings did not have any significant correlation with errors and with the output of UX_Mate.

Discussion

This pilot reports a preliminary evaluation of UX_Mate on the evaluation of 4 media players triangulating different measurement techniques. Overall, all measures (performance data, questionnaire scores and facial motion features) could consistently identify the worst system (i.e., MusicBee), although task variability was detected only by performance data and not by facial motion features. We found evidence of correlations between facial expressions as detected by UX_Mate and the number of errors. In particular, the regression analysis suggested that movements of the eyebrows were a powerful predictor of error occurrence, whereas mouth expressions were weaker and could not discriminate the worst system if used in isolation. This effect may be due to the tendency of participants of touching the lower part of their face, particularly in the moment of tension leading to video occlusion and hampering automatic detection of their emotional state.

The human-validation study suggested that UX_Mate can perform a task which is difficult and extremely expensive for human evaluators. The inference of behavioral cues based on facial cues requires a high level of specialization, as highlighted by the differential performance between trained and untrained observers. Overall, untrained observers appeared to express random evaluation with no correlation with each other, actual behaviour, or facial movements as highlighted by UX_Mate.

Overall, facial expressions showed a strong inter-individual variability within the sample. Some users had a very expressive and clear facial vocabulary, while other users had almost no apparent variations. This individual difference suggested the need of training UX_Mate to recognize the user facial expressivity before the evaluation task via a set of calibration tasks. Such tasks need to be designed carefully to identify the behaviour of interest, and need to be extremely short not to disturb the evaluation process. In the validation study, we introduced two calibration tasks in order to gain knowledge about a participant’s facial behavior in situations of different complexity.

The study also highlighted the limitations of post-test questionnaires in assessing the dynamic of experience evolution over the time course of the interaction. Indeed, the total number of errors and the subjective evaluations at the UX dimensions were only loosely correlated. On the contrary, we found evidence of a peak-end experience effect [20]: participants’ evaluation tended to reflect their performance in the last task which was significantly more complex than the previous one. This result is consistent with a growing body of research, showing that when people

construct summary judgements they are not only influenced by the average or the sum of their experiences, but that the final episodes and their valence have a major influence on summary judgement [1].

VALIDATION STUDY

A validation experiment was conducted on a set of videos collected during the evaluation of three social network services directed at people who study or work in universities or research institutions. The objective of the study was to understand how well UX_Mate predicts task difficulty through automated analysis of video recordings of the face of the users. The performance of UX_Mate was calculated by comparing the system's prediction of error occurrence against expert-based annotation of the errors performed by participants while completing the tasks they were assigned. Participants' faces were recorded by the webcam of their laptops and the actions performed on the interface were captured. Webcam recordings served as input for UX_Mate and for training the system to predict usability issues on the basis of facial emotional cues. Screen captures were used for manual annotation of errors. Moreover, participant evaluation of tasks and websites was collected during the experiment by means of questionnaires and compared to system predictions and expert-based evaluation. The experiment followed a similar procedure as the pilot study thus, only variations are reported.

Method

Participants

Thirty-one students (26 M, mean age = 28.5 years) of a local University were involved on a voluntary basis as participants in the study. Almost all the participants declared to have experience in using social network sites except two (1M, 1F). None of them ever used the sites tested before the study.

Procedure

The first step of the study consisted in the completion of a questionnaire assessing the participant's previous experience with academic social networks. Then, UX_Mate was calibrated by means of a task in which participants were asked to retrieve information in two conditions: *easy*, the target information was presented within a well-formatted table; and *difficult*, the target information was hidden among irrelevant information and visual clutter [35]. In the easy condition, which was taken as baseline, minimum modifications of facial expression were expected, whereas in the difficult condition the nature of the task was expected to elicit changes in facial expression connected to task difficulty.

After calibration was completed, participants were provided with the links to the homepages of the three academic social networks (academia.edu, researchgate.net and iamresearcher.com) and asked to perform the following tasks on each of them: 1) create an account; 2) edit profile by uploading a photo; 3) search for a publication 4) delete the account. The websites were presented in a random order, while the order of tasks was the same for the three websites. After each task, the participants were asked to rate its difficulty on a 7-point Likert scale. After interaction with each website, users filled in the UX questionnaire.

Video analysis

A detailed analysis was performed on the interaction video of each participant, annotating starting and ending point of every error[2]. Following the operational definition adopted in the pilot study, all variations from optimal performance were scored as errors. Double coding was performed on the entire sample and discrepancies were solved by discussion. A *performance* index was finally computed by dividing the number of video-frames spent in error and the number of total video-frames per task.

Results

Calibration task

A set of t-test was run to compare the activation level of the 4 MUs correlated to task difficulty and of the compound variables between the two calibration conditions. Bonferroni corrections were applied. All features analysed evinced a significant higher activation in the complex condition than in the easy one (Table 5).

	Simple task		Difficult task		t-test
	Mean	Std. error	Mean	Std. error	
MU4	65.23	20.72	188.65	41.26	-4.009***
MU6	85.87	22.32	182.77	43.92	-2.245*
MU7	44.39	9.85	171.26	40.49	-3.135***
MU8	79.42	21.3	227.23	58.29	-3.105**
Frustration left side	66.77	18.5	227.87	45.11	-2.606*
Frustration right side	79.35	20.32	229.9	61.87	-2.573*
Confusion	64.87	20.32	238.03	62.85	-2.430*
Overall activity	92.26	17	220.06	63.85	-2.098*

Table 5. t-test analysis * p < .05; ** p < .01 *** p < .001

Performance data

A sample of 358 evaluation tasks was collected and analysed. The average number of errors per task was 0.86, ranging from a minimum of 0 to a maximum of 7. Some 49% of the tasks contained no error, 39% of them contained 1 or 2 errors, 8% contained 3 or 4 errors, and the remaining 3% contained 5 or more errors. On the average, participants spent 13% of the interaction time performing wrong actions.

The *performance* index was analyzed by an ANOVA with

[2] The annotated corpus is available at disi.unitn.it/~staiano/trentoux

website (3) and tasks (4) as the within-subjects factors. The effect of task was significant ($F_{(3,75)} = 10.54$, $p < .001$, $\eta_p^2 = .3$). Post-hoc analysis showed that during task 1, participants spent less time performing wrong actions than during the other tasks. No other significant effects were returned.

	Task 1	Task 2	Task 3	Task 4
Usability	-.21	-.32	-.52**	-.31
Classical Aesthetics	-.17	-.32	-.50**	-.3
Expressive Aesthetics	-.18	-.3	-.50**	-.31
Summary Judgement	-.19	-.35	-.52**	-.19

Table 6. Correlation matrix * p < .05; ** p < .01

Performance/questionnaire correlation

All UX dimensions tested in the study reported high reliability values at the Cronbach test ($\alpha > .82$). Average scores were computed and used in the following analyses. The correlations between UX dimensions and perceived task difficulty indicated significant negative correlation only for one of the four tasks performed in the study (Table 6). For the other tasks, no associations were found.

	Task 1	Task 2	Task 3	Task 4
Website 1	.01	.14	.24	.51**
Website 2	.50**	.50**	.40*	.22
Website 3	.04	.40*	.46*	.10

Table 7. Correlation matrix * p < .05

The correlation matrix between perceived task difficulty and the performance index is reported in Table 7. The significant coefficients show a positive correlation between perceived difficulty and performance index only for half of the tasks analysed.

The correlation between the percentage of task spent in error (total and for the 4 tasks separately), and the 5 UX dimensions highlights the same peak-end experience effect as in the pilot study. Significant correlations appeared only with respect to the performance to the last task.

Automatic UX assessment

In order to assess the performance of UX_Mate, a bayesian model was trained using the Activation-Indexed data resulting from the calibration tasks and evaluated using the Activation-Indexed data resulting from the evaluation tasks. The model returned a set of predictions indicating if a task was simple or difficult. These results were compared against the expert-based annotations of the evaluation tasks. Tasks were marked as simple if they contained no errors, otherwise they were marked as difficult.

The confusion matrix, reported in Table 8, visualizes UX_Mate's performance in predicting simple or difficult tasks based on the presence of usability errors (annotated during the video analysis). The highest the values along the diagonal, the better the system's performance. It is evident that UX_Mate performed reasonably well: it correctly predicted the class of more than two thirds of the tasks.

Task	Simple	Difficult
Simple	101	71
Difficult	35	151

Table 8. Confusion matrix.

In particular, it was able to correctly classify 101 tasks as simple, out of the 136 tasks marked as simple in the dataset (column 1). Conversely, it was capable to identify 151 tasks as difficult out of the 222 marked as difficult in the data set (column 2).

Table 9 summarises the experimental results considering three standard performance indexes used in machine learning. Precision is the number of tasks identified correctly out of the total number of tasks. Recall is the number of correct results returned by the model, divided by the number of results that should have been returned. The F1-measure is considered as the ultimate quality metrics of a classifier in machine learning. It provides a measure of the test accuracy in a two-class binary classification. The F1-Measure is defined as the harmonic mean of precision and recall, therefore its values provides an estimate of both variables. Overall, UX_Mate obtained a slightly higher precision in classifying simple tasks than in classifying difficult ones. On the other hand, a much larger difference emerged with recall: 81% of the tasks marked as difficult in the dataset were correctly identified as such.

Task	Precision	Recall	F1
Simple	.74	.59	.66
Difficult	.68	.81	.74

Table 9. UX_Mate's experimental results.

Discussion

The study provided encouraging evidence on the validity of UX_Mate as a tool to predict the occurrence of usability errors. Overall, the system was capable to discriminate between tasks containing and tasks non-containing errors in over two thirds of the cases. The study also confirmed the complexities of UX evaluation as assessed by questionnaire data. Not only the summary judgement was strongly biased towards the performance of the final tasks, as previously highlighted in the pilot study, but also the pattern of association between different types of subjective data collected at different stages of the evaluation was complex. Significant correlations between perceived difficulty and UX dimensions were systematically found only for one specific task. Finally, we found weak evidence of correlations between participants self-reports on task

difficulty and occurrence of usability errors, showing that these two concepts may not always be related.

We believe that the performance of UX_Mate can be improved by collecting a larger database of appropriate calibration tasks on which to train the model used by the machine learning module (Figure 1). It is fundamental that these calibration tasks are unambiguously linked to the interaction events extracted during the prediction phase. This aspect was a limitation of our study. Indeed, the calibration task we used, despite having the advantage of being extremely short and already tested in the literature [35], had only an indirect link with the event of interest implied by the equation: usability problems = increased difficulty. The mismatch was evident in the questionnaire results. Further improvement to the performance of UX_Mate can be achieved by setting more discriminative thresholds between the classes predicted by the machine learning module. The validation study reported in this paper addressed professionally designed websites, with less than 1 error per average task. As a consequence, we had to use two broad prediction classes, just considering presence versus absence of errors. Despite this sampling limitation, UX_Mate was still able to recognize the occurrence of interaction errors as tagged by expert evaluators even when the task was not perceived as difficult by the users.

CONCLUSION

This study presented UX_Mate, a modular system which tracks facial expressions of the users, interprets them based on pre-set rules, and generates predictions about the occurrence of target behaviour during HCI. Prediction is based on facial expression examples collected in the calibration phase. In this paper, we concentrated on prediction of errors occurrence from facial expressions linked to frustration and confusion. In future research, we aim to extend this paradigm to other interaction feelings, such as enjoyment or boredom, in order to study what interaction features may be responsible for their occurrence. UX_Mate is designed in a modular way, allowing the evaluator to choose the machine learning algorithms, and the set of examples to train the system that best fit their particular needs. We have now collected a new large database of facial and interaction videos in an entertainment setting focused on feelings like flow, engagement and fun.

UX_Mate represents a preliminary yet important step towards the automatic assessment of User eXperience. We believe that the automatic assessment of facial expressions can be a powerful tool to support UX studies following a research paradigm based on the triangulation of different techniques, including human-based observation, self-reports, and facial expression tracking. In particular, UX_Mates can provide a cheap method to monitor the dynamic evolution of emotions in time, and counteract the tendency of questionnaires to anchor on specific moments of the performance. An automatic tool can be an important help to untrained evaluators when they need to understand emotional reaction. UX_Mate can perform a task which is difficult and extremely expensive for human evaluators. Indeed the pilot study demonstrated that inferring performance behaviour based on visual cues from the participant face is time consuming and requires a high level of specialization.

Contrary to [37], we claim that UX research requires methods that can be applied beyond the laboratory settings of a usability laboratory. In this respect UX_Mate represents a unique and promising tool. The data collected during the Validation Study will be contributed to the community, along with the corresponding expert annotations and guidelines to further extend the dataset. Our hope is that such a contribution will boost and facilitate the research on automatic methods for UX assessment.

ACKNOWLEDGMENTS

We are grateful to Zahid Hasan and Suresh Daggumati for their help with data collection and coding. A special thank to Silvia Torsi for her contribution and interesting feedback, and to all participants for their time and faces. The work of J. Staiano and N. Sebe has been partially supported by the FIRB project S-PATTERNS.

REFERENCES

1. Ariely, D. and Carmon, Z. Gestalt characteristics of experiences: the defining feature of summarized events. Journal of Behavioural Decision Making (2000), 13 (2), 191-201.
2. Benedek, J., and Miner, T. Measuring desirability: New methods for evaluating desirability in a usability lab setting. UPA (2002).
3. Bradley, M., and Lang, P.J. Measuring emotion : The self-assessment semantic differential manikin and the semantic differential. Science 25, I (1994), 49-59.
4. Branco, P., Firth, P., Encarnação, L.M., and Bonato, P. Faces of emotion in human-computer interaction. In CHI EA (2005), 1236-1239.
5. Burr, B. Vaca: a tool for qualitative video analysis. In CHI EA ’06 (2006), 622-627.
6. Chorianopoulos, K., and Spinellis, D. User interface evaluation of interactive tv: a media studies perspective. Univers. Access Inf. Soc. 5 (2006), 209-218.
7. Cohen, I., Sebe, N., Chen, L., Garg, A., and Huang, T. S. Facial expression recognition from video sequences: Temporal and static modelling. In Comput. Vis. Imag. Und. (2003), 160-187.
8. Dacher, K. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. Pers. Soc. Psychol. 68, 3 (1995), 441-454.
9. Davidson, R.J., Ekman, P., Saron, C.D., Senulis, J.A., and Friesen, W.V. Approach-withdrawal and cerebral asymmetry: Emotional expression and brain physiology.

Pers. Soc. Psychol. 58, 2 (1990), 330-341.

10. Den Uyl, M., Van Kuilenburg, H., and Lebert, E. Facereader: an online facial expression recognition system. In Measuring Behavior (2005).
11. Derbaix, C.M. The impact of affective reactions on attitudes toward the advertisement and the brand: A step toward ecological validity. J. Marketing Res. 32, 4 (1995), 470-479.
12. Desmet, P. Measuring emotion: development and application of an instrument to measure emotional responses to products. Kluwer Academic Publishers (2004), 111-123.
13. Ekman, P. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. W. W. Norton & Company, (1992).
14. Ekman, P., and Friesen, W. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, (1978).
15. Grootjen, M., Neerincx, M.A., van Weert, J.C.M., and Truong, K.P. Measuring cognitive task load on a naval ship: implications of a real world environment. In Proc. FAC (2007), 147-156.
16. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. The weka data mining software: an update. SIGKDD Explor. Newsl. 11, (2009), 10-18.
17. Hassenzahl, M., and Tractinsky, N. User experience - a research agenda. Behav. Inf. Technol. 25, 2 (2006), 91-97.
18. Hazlett, R.L., and Benedek, J. Measuring emotional valence to understand the user's experience of software. Int. J. Hum.-Comput. Stud. 65 (2007), 306-314.
19. Isbister, K., Hk, K., Laaksolahti, J., and Sharp, M. The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect. Int. J. Hum.-Comput. Stud. 65, 4 (2007), 315-328.
20. Kahneman, D., Fredrickson, B. L., Schreiber, C. A. and Redelmeier, D. A. When more pain is preferred to less: Adding a better end. Psychological Science , 4 (1993), 401-405.
21. Kuilenburg, H.V., Wiering, M., and Uyl, M.D. A model based method for automatic facial expression recognition. In Proc. ECML (2005), 194-205.
22. Lavie, T. and Tractinsky, N.: Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. Int. J. Hum.-Comput. Stud. 60, 3 (2004), 269-298.
23. Lopatovska, I., and Arapakis, I. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. Inf. Process. Manage. 47 (2011), 575-592.
24. Mahlke, S., Minge, M., and Thüring, M. Measuring multiple components of emotions in interactive contexts. In CHI EA '06 (2006), 1061-1066.
25. Mcdaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., and Graesser, A. Facial Features for Affective State Detection in Learning Environments. In Proc. 29th Annual Meeting of the Cognitive Science Society (2007).
26. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Curr. Psychol. 14, 4 (1996), 261-292.
27. Pantic, M., and Rothkrantz, L. Automatic Analysis of Facial Expressions: The State of the Art. IEEE T. Pattern Anal. 22, 12 (2000), 1424-1445.
28. Picard, R.W. Affective Computing for HCI. In Proc. HCI 1999, 1, 1 (1999), 829-833.
29. Ravaja, N., Turpeinen, M., Saari, T., Puttonen, S., and Keltikangas-Järvinen, L. The psychophysiology of James Bond: Phasic emotional responses to violent video game events. Emotion 8, 1 (2008), 114-120.
30. Ryan, A., Cohn, J., Lucey, S., Saragih, J., Lucey, P., De la Torre, F., and Rossi, A. Automated facial expression recognition system. In Proc. IEEE ICCST (2009), 172-177.
31. Scheirer, J., Fernandez, R., Klein, J., and Picard, R.W. Frustrating the user on purpose: A step toward building an affective computer. Interact. Comput. (2001) 93-118.
32. Sebe, N., Lew, M.S., Cohen, I., Sun, Y., Gevers, T., and Huang, T. S. Authentic facial expression analysis. In Proc. IEEE AFGR (2004), 517-522.
33. Swindells, C., Maclean, K.E., Booth, K. S., and Meitner, M. A case-study of affect measurement tools for physical user interface design. In Proc. GI (2006), 243-250.
34. Tractinsky, N. and Zmiri, D.: Exploring attributes of skins as potential antecedents of emotion in HCI. Aesthetic computing (2006), 405-422
35. Tullis, T.S. Screen design. Handbook of Human-Computer Interaction (1997).
36. Westerman, S. J., Sutherland, E. J., Robinson, L., Powell, H., and Tuck, G. A multi-method approach to the assessment of web page designs. In Proc. ACII (2007), 302-313.
37. Zaman, B., and Shrimpton-Smith, T. The facereader: measuring instant fun of use. In Proc., NordiCHI '06 (2006), 457-460.
38. Joho, H., Staiano, J., Sebe, N., and Jose, J. Looking at the Viewer: Exploiting Facial Activities to Detect Personal Highlights of Multimedia Contents. Int. J. Multimed. Tools App. 51, 2 (2011), 505-523.