UX Heatmaps: Mapping User Experience on Visual Interfaces

Vanessa Georges
HEC Montréal
Montréal, Canada
vanessa.georges@hec.ca

François Courtemanche
HEC Montréal
Montréal, Canada
francois.courtemanche@hec.ca

Sylvain Sénécal
HEC Montréal
Montréal, Canada
ss@hec.ca

Thierry Baccino
Université de Paris VIII
Paris, France
baccino@lutin-userlab.fr

Marc Fredette
HEC Montréal
Montréal, Canada
marc.fredette@hec.ca

Pierre-Majorique Léger
HEC Montréal
Montréal, Canada
pml@hec.ca

ABSTRACT
In this paper, we present an off-the-shelf UX evaluation tool which contextualizes users' physiological and behavioral signals while interacting with a system. The proposed tool triangulates users' gaze data with inferred users' cognitive and emotional states to produce user experience (UX) heatmaps, which show where users were looking when they experienced specific cognitive and emotional states. Results show that for a given cognitive state (i.e., cognitive load), the proposed UX heatmap was able to effectively highlight the areas where users experienced different levels of cognitive load on an interface. The proposed tool enables the visual analysis of users' various emotional and cognitive states for specific areas on a given interface, and also to compare users' states across multiple interfaces, which should be useful for both UX researchers and practitioners.

Author Keywords
User experience; interface design; heatmaps; eye tracking; physiological computing; cognitive load; affective computing.

ACM Classification Keywords
H.2.1Design; Experimentation; HCI; Human Factors; Measurement; User interfaces.

INTRODUCTION
High quality user experience (UX) has become a key competitive factor for product development [17]. User experience methods investigate how people feel about a system, game, or web interface, as opposed to how easily they managed to accomplish the task at hand, shifting the focus towards user affect. Assessing the range and quality of affects experienced while interacting with a system is now seen as critical for the development of products that satisfy both users' needs and expectations [14]. Measuring the emotional state of users during the interaction is therefore essential to the design of richer user experience. Traditional evaluation methods (e.g. questionnaires and interviews) rely on self-reported data in order to assess the affective and cognitive states of users, which are often exposed to different response effects, such as social desirability [18]. More so, these methods assess user perceptions either after or during the interaction, which induces retrospective biases or disrupts the user experience [9, 27].

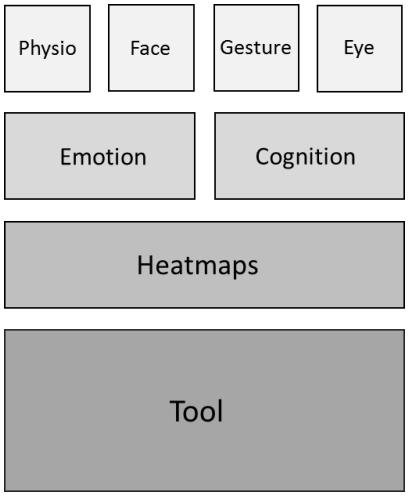

Figure 1. From signals to UX heatmap tool

Users' emotional and cognitive states can also be inferred using many different types of physiological and behavioral signals, such as electrodermal activity, heart rate, eyetracking, vocal and visual cues, body gestures, or facial expressions (see [37] and [6] for reviews). The dynamic nature of these signals offers a continuous window to the users' reactions, and can provide valuable insights as to what they are experiencing during the interaction. These signals

CHI'16, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-3362-7/16/05.
DOI: http://dx.doi.org/10.1145/2858036.2858271

can also be used to infer emotional and cognitive states, some of which the user himself is either unaware of or cannot recall when asked using traditional deferred methods [34]. Although physiological and behavioral measures are used in academia, and to a lesser extent in industrial contexts, many challenges remain [32]. These methods are often costly and require expert knowledge. However, the main obstacle to the use of physiological and behavioral signals remains their reduced informative value when they are not specifically associated with user behavior or interaction states [11, 25].

To meet this challenge, most researchers have concentrated their efforts on finding ways to measure physiological signals and interaction states synchronously. For example, Kivikangas et al. [19] have developed a triangulation system to interpret physiological data from video game events. Dufresne et al. [10] have proposed an integrated approach to eyetracking-based task recognition as well as physiological measures in the context of user experience research. Other researchers have also developed tools that allow users' to manually assign subjective emotional ratings on visual interfaces [16] or to visualize emotional reactions in terms of GUI widgets [7]. While these research streams have produced interesting results, they are not easily transferable to new contexts of use, as they are based on internal information from the interactive system (e.g., video game logs, application events, or areas of interest).

This paper presents a novel off-the-shelf and easy to interpret UX evaluation tool that aims to contextualize users' signals while interacting with a system interface. As illustrated in Figure 1, different emotional (sadness, happiness, surprise, etc.) and cognitive (cognitive load, stress, etc.) states are first inferred from continuous physiological or behavioral signals. The states are then triangulated with gaze data and mapped onto the user interface in order to create heatmaps and highlight areas where they occur with a higher frequency. The tool operationalizes this method by implementing different types of heatmap visualizations and evaluation functionalities. In order to describe the proposed tool, this paper focuses on the visualization of users' psychological states using physiological signals. Cognitive load is used as the case example.

The remainder of the paper is organized as follow. First, a detailed account of the creation of physiological heatmaps is followed by an overview of the proposed tool. Experimental validation results are presented. Applications of the tool to UX evaluation is discussed in the form of a use case using experimental data. Finally, the genericity of the tool is illustrated with the use of emotional heatmaps based on facial expressions.

PHYSIOLOGICAL HEATMAPS

Traditional gaze heatmaps are used in eyetracking as intuitive representations of aggregated gaze data [22]. Their main use is to help researchers and HCI experts answer the question: "Where in the interface do people tend to look?" [35]. In the proposed visualization method, the users' gaze now serves as a mean of mapping physiological signals onto the user interface. The resulting heatmaps represent the physiological signals' distribution over the interface, and can help answer the following question: "Where in the interface do people tend to emotionally or cognitively react more strongly?" Below, we describe the four steps involved in the creation of physiological heatmaps: inference, normalization, accumulation, and colorization.

Inference

The regulation of emotional and cognitive states relies at once upon the sympathetic and parasympathetic activity of the autonomic nervous system, and thus requires physiological adjustments stemming from multiple response patterns [20]. As such, the relation between physiology and psychological states is more realistically described as a many-to-many relationship (i.e., multiple psychological states linked to multiple physiological variables) [5]. For example, when interacting with an interface, a single user's physiological signal (e.g., heart rate) can be associated with a change in cognitive load or emotional arousal. Therefore, the objective of the first step of the creation process is to disentangle this effect and ensure that the rendered heatmap truly represent the psychological construct of interest (i.e., cognitive load in the context of this paper). To do so, a machine learning classifier is used to estimate users' cognitive load.

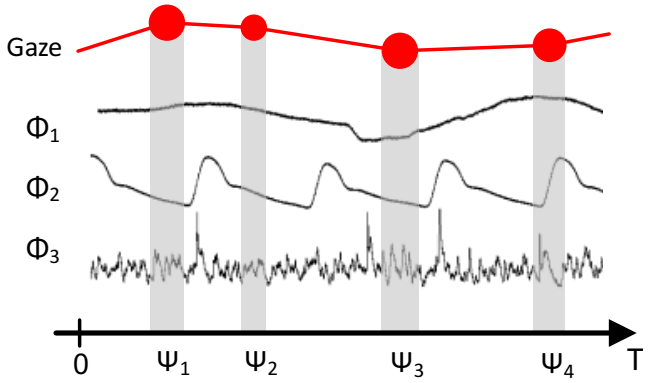

Figure 2. Gaze-based psychophysiological inference process. Red dots represent eye fixations and red lines represent saccades. Φ1 = electrodermal activity, Φ2, = blood volume pressure, and Φ3 = pupil size. Ψ_x represent different inferences of a given construct of interest (e.g., different inferences of cognitive load).

As illustrated in Figure 2, all signals (Φ_1, Φ_2, Φ_3) are segmented according to users' gazes and serve as features in the classifier. As each physiological system operates in collaboration with a variety of inputs and outputs from the rest of the organism, the measured signals present various durations and latencies for a given stimulus. For example, heart rate may have a shorter latency than electrodermal activity for a given stimulus. Therefore, each signal is segmented using a specific extraction window of different duration and latency starting at fixation onset. Latency is defined as the time elapsed between a fixation onset and the

beginning of the extraction window and duration is defined as the time elapsed between the start and the end of the window. The identification of optimal latencies and durations, as well as the training of the classifiers are done using an empirical optimization process. For this study, the windows' optimization was done using a subpart of the data collected in the experiment. A detailed description of the segmentation and machine learning processes is given in [8]. This paper focuses on the heatmap visualization aspect of the tool. For each user's fixation, the inference outcome (Ψ_x) is used in the next step of the creation process.

Normalization

As physiological signals are subject to significant interpersonal variations or instrumental inaccuracies, absolute values cannot be used to compare data from multiple users. Physiological signals need to be corrected to account for the user's baseline [31]. In the proposed approach, the results of the physiological inference (see inference section) are normalized using z-score with the following equation:

$$W'_i = \frac{W_i - \mu}{\sigma}$$

where μ and σ are respectively the mean and the standard deviation of the inferred values for all of a user's fixations. This step also helps distinguish the "physiologically significant" areas of an interface from neutral ones. In traditional gaze heatmaps, every fixation has a positive intensity value and increases the height map. In our approach, physiologically unimportant fixations ($W' < 0$) are not considered in the subsequent accumulation step; only important physiological activity makes a contribution.

Accumulation

For traditional gaze heatmaps, the accumulation step consists in the creation of a blank map with the same dimensions as the image stimulus (n x m pixels). For each eye fixation, all the pixels are attributed an intensity level corresponding to an eyetracking metric (fixation count, absolute fixation duration, relative fixation duration, or participant percentage) multiplied by a scaling function (e.g., Linear, Gaussian), taking into account the distance between the gaze and the pixel [15]. Intensity values from various users falling on the same pixel are then summed to produce a height map representing the gaze intensity distribution over the image stimulus (Figure 3).

In the proposed method, the inference outcome (Ψ_x) is used to create the aforementioned height map, instead of an eyetracking metric. Therefore, the resulting height map represents the relative intensity of the physiologically inferred psychological state over the interface. Furthermore, as for standard gaze heatmaps, the height map is not rendered in its entirety. In order to outline the most prominent parts of an interface, a parameter $t \in [0, 1]$ is defined to compute the threshold under which intensity values are neglected. For example, $t = 0.2$ implies that only pixels with an accumulated intensity superior to 20% of the maximum intensity will be rendered in the colorization step. The value of the t threshold can be seen as representing the water level in a flooded valley. In this study, heatmaps were generated using Gaussian scaling and an intensity threshold of 0.8.

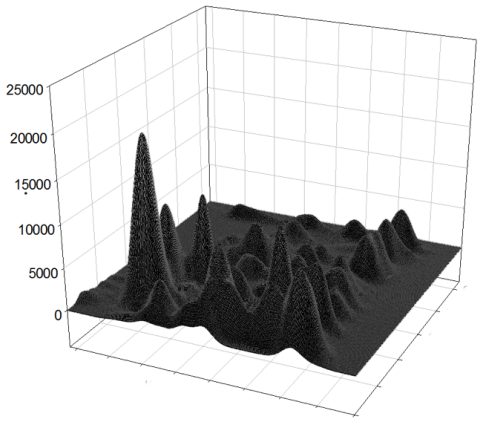

Figure 3. Height map representation of aggregated gaze data.

Colorization

The last step in creating a heatmap is colorization. The main idea is to overlay on the stimulus image a semi-transparent layer that reflects the height map's variations. Height variations can be mapped to different color properties using a colorization function, resulting in various types of visualizations. The most commonly used visualizations are rainbow heatmaps, luminance maps, and contrast maps [15]. Breslow et al. [4] have demonstrated that multicolored scales (e.g., rainbow) are best suited for identification tasks (i.e., determining absolute values using a legend) and single hue scales (e.g., luminance) are best suited to compare relative values. Therefore, in this work, the colorization step uses a luminance gradient (i.e., an increase in color brightness is associated with an increase in the height map). Furthermore, the use of single hue gradients allows the mapping of multiple constructs at the same time.

Cross-stimuli Colorization

The proposed tool also features a novel colorization function: cross-stimuli gradient. This colorization function displays a single gradient across multiple stimuli, in order to evaluate users' experience of one interface relative to others.

Eyetracking analysis software provide heatmap representations on a single image basis only. Heatmaps are generated one at a time and the full gradient (i.e., all possible color properties increments) is displayed on each image. The colorization function associates pixels under the maximum height with the maximum value of the color gradient, and pixels at the threshold height with the minimum value. For example, when generating a rainbow heatmap on a given interface, a gradient going from red to green, by way of yellow and orange, is mapped in its entirety onto the image. In a luminance gradient, the maximum value of the color gradient is the absence of hue, i.e. white. This

implementation is relevant for gaze heatmaps as they are meant to represent how a specific amount of aggregated eye fixations is distributed within a single interface. However, it may be interesting to evaluate the user's experience of said interface relatively to different interfaces.

To do so, we propose a colorization function that uses a cross-stimuli gradient. First, the height maps of a specific set of image stimuli that are to be analyzed together are merged together. Second, the colorization function associate pixels' height of each rendered heatmap to a gradient level, relatively to its location in the common height map. Therefore, the full gradient representing the user experience construct spans across all images and, allows users to compare areas of different interfaces together. Unlike standard gaze heatmaps, heatmaps rendered using this cross-stimuli gradient do not necessarily present hot spots (i.e., doesn't use the full gradient within the same interface). Hot spots are present only on interfaces that present the highest density of the visualized user's state compared to the others. For example, figure 5 illustrates three interfaces representing various levels of visual complexity (low, medium and high). Instead of comparing three individually constructed heatmaps (i.e. one per image), a single cross-stimuli gradient is mapped across all three stimuli. The hot spot is located on the top left corner of the third stimulus, meaning that users experienced the highest level of cognitive load in this specific area, relative to the other versions of the interface.

UX HEATMAP TOOL DESCRIPTION

The evaluation tool enables the contextualization of various physiologically inferred affective and cognitive states, during users' interaction with an interface. Figure 4 illustrates the data processing sequence required to generate heatmaps.

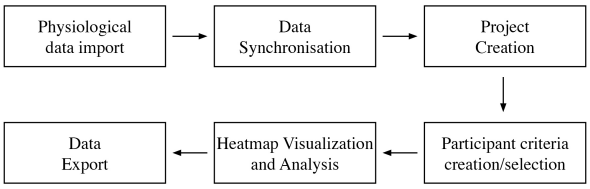

Figure 4. Data process sequence

First, data files from various physiological and behavioral recording devices are imported and synchronized. Upon project creation, participant selection (e.g., women only) is required, followed by heatmap creation and rendering. Multiple heatmaps can be displayed simultaneously (e.g., negative valence and cognitive load, or negative and positive valence). The opacity and display color of each heatmap can be individually selected, before or after visualization, in order to enable the comparison of multiple emotional and cognitive constructs. The resulting heatmaps can then be exported in jpeg format.

Participant Selection

Participant data can be created and managed before or after the creation of a new project. This data consists of an identifier and variables, such as gender and age. These can be used to create participant groups; which users can subsequently use to filter the physiological data when producing heatmaps. These variables can be selected using a

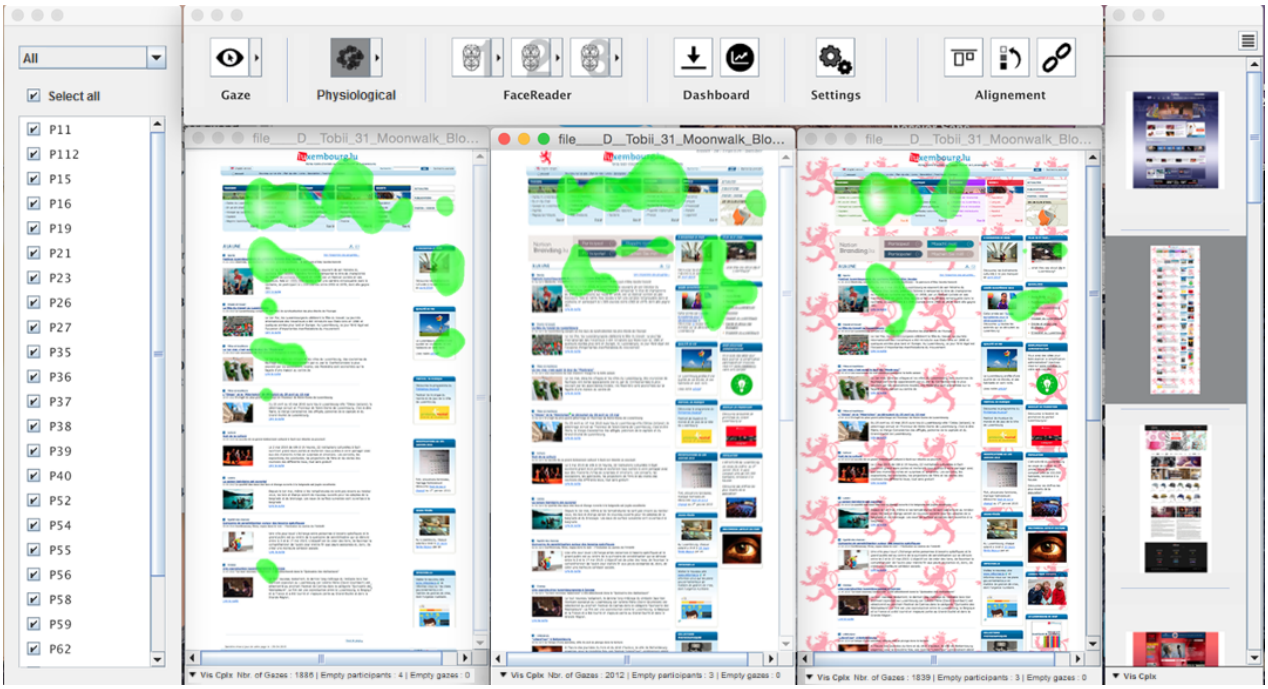

Figure 5. Participant data can be created and managed using the left hand panel. Images used during a test or experiment are displayed on the right side panel. Above: An example of a cross-stimuli gradient is illustrated using images representing three levels of cognitive load (from left to right: low, medium and high).

drop-down menu on the left hand panel. All participants are displayed in the participant list (see Figure 5).

Construct Manager

Three types of heatmaps can be generated in the current state of the tool: gaze heatmaps, physiological heatmaps (based on cognitive load), and facial expressions heatmaps (emotion). Emotion heatmaps are generated using the FaceReader 6 software (Noldus, Netherland), which infers the probability of seven discrete emotions (happy, sad, angry, surprised, scared, disgusted and neutral) and emotional valence (negative vs. positive), based on facial movements [36]. Given that FaceReader recordings are synchronized with eyetracking data, emotion prediction is introduced in the accumulation step of heatmap creation (see physiological heatmaps section). As seen in Figure 5, up to three emotions can be mapped onto the interface at any given time during the analysis, to allow the simultaneous comparison of various emotional states. Heatmaps can all be displayed one by one or simultaneously based on user selection. Participant selection and group filters are applied to all visualizations.

Interface Selection

Images used during the analysis are displayed on the right side panel. Users can select as many stimuli as desired. Upon heatmap visualization, information relative to selected stimuli (e.g., number of gazes) will appear at the bottom of the prompted window.

EXPERIMENTAL VALIDATION

A lab experiment was conducted to develop and evaluate the proposed tool. In this validation study, cognitive load was used as the psychological construct of interest. The goal of the evaluation was twofold: 1) assess the ability of the physiological heatmaps method to effectively highlight the different levels of cognitive load experienced on an interface, and 2) present a use case in order to illustrate the usefulness of the proposed tool for user experience evaluation. Two separate tasks were designed to achieve these goals.

Participants

For this experiment, a total of 44 students between the ages of 18 and 35 were recruited through the student panel at HEC Montréal over a period of five weeks. Data from 18 participants were rejected due to equipment malfunction, data synchronization imprecision, or insufficient eyetracking calibration precision. Therefore, data from 26 participants were used in the analyses, from which 17 were female, for an average age of 24. Participants had normal or corrected-to-normal vision and were pre-screened for glasses, laser eye surgery, astigmatism, epilepsy, and neurological and psychiatric diagnoses. The total experiment duration was of one hour, during which participants were asked to perform the two experimental tasks. Compensation in the form of a 20$ gift certificate was given to each participant upon completion of the experiment.

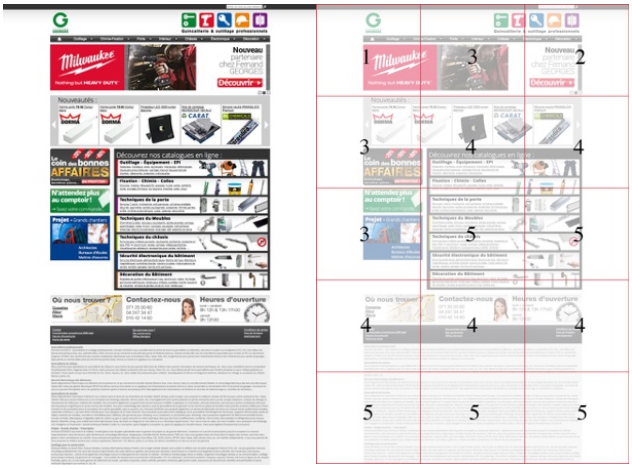

Figure 6. Left: Stimulus 7 was rated as medium complexity website, with inter-judged average visual complexity rating of 4.85 out of 10. The task pertaining to this particular stimulus, given to evaluating judges and subjects alike was to search for information regarding parts and equipment brands available in store. Right: Evaluation sheet for Stimulus 7 with an example of a participant's ratings.

Stimuli

In the experiment, the visual complexity of website homepages was manipulated in order to trigger different cognitive load levels. Visual complexity is closely related to the perceived cognitive load users think will be required to interact with an interface [13] and has also been shown to be correlated with cognitive load [33]. Multiple studies have aimed at understanding how the cognitive system interprets different levels of image complexity (see [37], [6] and [26]).

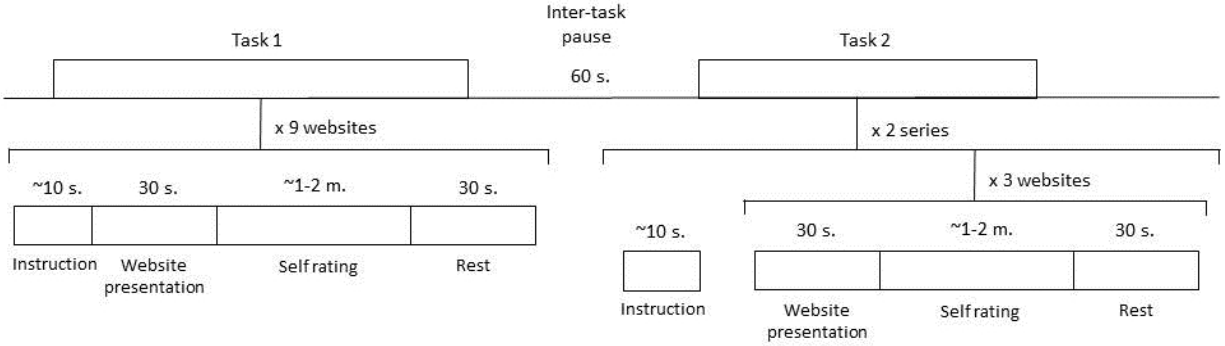

Figure 7. Experimental procedure

Most studies consider that there are three types of cognitive load (CL): intrinsic load, extraneous load, and germane load. According to Sweller [30], intrinsic cognitive load is linked to the material's content, extraneous cognitive load is based on the presentation forms, and germane cognitive load involves information consolidation. Website complexity relates to extraneous cognitive load [31]. Wang et al. [31] also found that extraneous CL can be decreased by adequate visual presentation and design of material.

These images were presented in an Internet browser, as to maintain the aspect and the original length of the homepage. Interaction with homepages was limited to scrolling. Websites were selected to reflect a wide range of tasks and sectors, such as business, academia, entertainment and leisure. In order to eliminate content-related affect, news websites were excluded. Local and national websites were also excluded in order to eliminate any bias based on familiarity.

Task 1

The goal of the first experimental task was to gather a physiological data set representative of different levels of experienced cognitive load, in order to assess the accuracy of the proposed method. The stimuli included nine homepages representing three levels of visual complexity: 3 low, 3 medium, and 3 high.

In order to select these nine homepages, twelve undergraduate students were instructed to rate the overall visual complexity of twenty-one website homepages, as defined as the degree of difficulty in understanding the information presented. The complexity scores' agreement was calculated using a two-by-two correlation procedure. Agreement between all 66 possible pairs of judges was averaged for each individual homepage. The resulting average correlation (for all homepages) was of $r=0.601$ ($p=0.047$). The stimuli were presented as illustrated in Figure 7 (Task 1).

At the beginning of each trial, participants were given a task to perform on the forthcoming homepage. They were instructed to look at the homepage having in mind how they would execute the task (e.g., search information about products and brands available in store). The website was then presented in a web browser for 30 seconds. Following the presentation, a printed version of the same interface was presented for an unlimited amount of time, featuring a red 15 box grid (see Figure 6). Participants were asked to assess the visual complexity of each of the 15 regions of the homepage using a 1 to 5 rating scale, 5 being very complex.

Task 2

The objective of the second task was to illustrate the usefulness of the proposed tool for webpage prototype comparison. Two series of three website homepages were doctored in order to reflect three different level of perceived visual complexity: low, medium, and high (Figure 5). Starting from a single homepage, two duplicate versions of the images were created. The interface was thus simplified (low complexity), as well as rendered more complex (high complexity). According to Oliva et al. [23], visual complexity is mainly represented by the perceived dimensions of quantity of objects, clutter, openness, symmetry, organization, and variety of colors. These elements were used to alter the complexity level of the images. The content was not altered. The images were evaluated by undergraduate students, in order to assess their visual complexity. They were modified and re- evaluated as need be. As illustrated in Figure 7 (Task 2), participants were first instructed on the task they would have to execute. The three versions of the website were then presented with a self-rating and rest period. The presentation order was randomized for each participant. The same procedure was repeated with another series of websites. Task 2 lasted twenty minutes on average.

Physiological Signals and Equipment

The signals used for physiological heatmaps should be selected according to the psychological construct of interest. Thus, signals known to be related to cognitive load were selected in this study.

A Biopac MP150 amplifier (Biopac MP) was used to record two peripheral physiological signals: electrocardiogram (ECG) and electrodermal activity (EDA). ECG-derived measures have been shown to correlate with affective and cognitive processes. For example, heart rate can be associated with shifts from low to high mental workload [12]. EDA measures the activity of the eccrine sweat glands and has been shown to be correlated to arousal. It can be used to measure emotions [3] and cognitive load [29] during system interactions. EDA was recorded using two electrodes placed on the palm of the non-dominant hand. Both physiological signals were recorded with a sampling rate of 500Hz.

Eye fixations, pupil diameter, and blink rates were recorded with a Tobii X-60 eyetracker (Tobii Technology AB). Research shows that variations in pupil size respond significantly to cognitive and emotional stimuli [21]. Ambient illumination was controlled and kept constant during the experiment in order to minimize pupil size variations due to light reflex. Blink rate can be affected by external stimulus or by emotional and cognitive states such as fatigue, but has been shown to be a perceptual load indicator (where information can be noticed or unnoticed) in the context of cognitive load measurement [1]. A nine point calibration was performed for all participants, and was repeated until sufficient accuracy was achieved. Current video-based eye trackers have a spatial resolution of up to 0.01^{o}/ 2 kHz [15]. However, such a high resolution level is experimentally hard to achieve due to subjects' variability. In this research, sufficient accuracy was defined as ~ 1 cm around the center of the calibration points, which led to participants' dismissal, if not obtained. Stimuli were

presented on a 22' LG LED monitor with a resolution of 1680 x 1050 pixels and a refreshing rate of 60Hz.

Videos of the participants' face were recorded using a webcam and the Media Recorder 2 software (Noldus, Netherland). Videos were processed in FaceReader 6 (Noldus, Netherland) to produce emotional states inference.

RESULTS

The effectiveness of the proposed method to predict spatial locations of greater cognitive load has been tested using data from Task 1. Data from Task 2 is analyzed qualitatively in the Discussion section. Data was analyzed in order to evaluate the capacity of physiological heatmaps to capture experienced cognitive load variance over the different webpages. As a relative baseline, correlation with gaze heatmaps was also calculated. Although they are not meant to assess cognitive load, traditional gaze heatmaps represent the closest data visualization method to which we can compare physiological heatmaps. As illustrated in Figure 8, data from Task 1 was analyzed by comparing the highest peaks of the intermediate height maps with the underlying user ratings. One height map was generated per participant for each homepage.

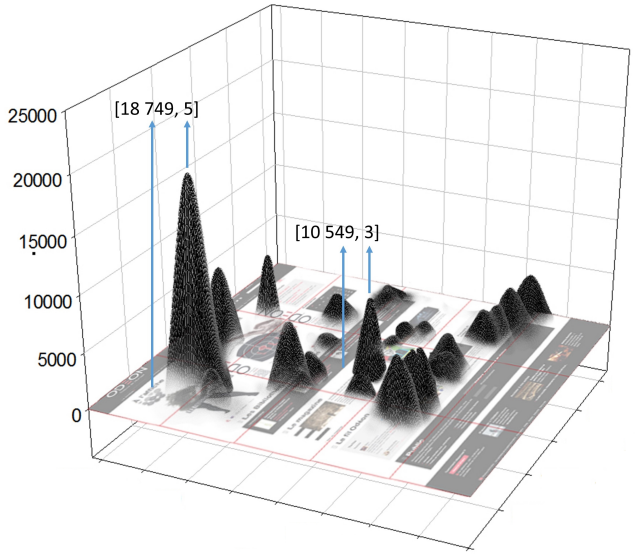

Figure 8. Evaluation data example. The first self-rating (top-left) square has a value of 5 (on a maximum of 5) and the highest peak of the overlapping portion of the height map is of 18 749 (relative scale). One data point is generated per rating square for a maximum of 15 per website per subject. The evaluation square over which the height map is null are not used in the analyses.

This process expected a total of 3 510 data points (9 websites x 15 ratings x 26 subjects). However, 2 749 data points were obtained for physiological heatmaps and 2 147 data points for gaze heatmaps as only ratings over which the height map is not null were considered. Therefore, websites' areas that were not looked at during the 30s presentation (but rated afterwards) were not used in the analyses. As shown in Table 1, physiological heatmaps were significantly related to users' ratings of visual complexity (R^2=.291, p-value<.000). Using the distribution of the R^2 statistics [24], we can also affirm that the R^2 obtained using the physiological heatmaps is significantly higher than the one obtained with gaze heatmaps (z=3.024; p-value=0.02).

Visualization	R^2 (p-value)
Physiological heatmaps	0.291 (<0.000)
Gaze heatmaps	0.167 (<0.000)

Table 1. R^2 between highest peaks and users' ratings

The p-values were corrected by assessing the effect of each measure on the ratings values by using linear regression model accounting for the potential correlation between each repeated measure coming from the same subject. These results show that cognitive load was successfully assessed by the proposed physiological heatmap, which can be complementary to traditional gaze heatmap. While gaze heatmaps can identify the "where", UX heamaps can be used to address the "why" of users' gaze behavior.

DISCUSSION

User experience evaluation needs differ between industry and academia. While the former's needs are to analyze and adequately communicate findings in order to improve UX, the latter's interest resides in the validation and understanding of phenomena, based on hypothesis [28]. This section illustrates two types of analysis that can be undertaken with the described tool: 1) a comparative analysis of various prototypes using a cross-stimuli gradient and, 2) an exploratory analysis using emotional and cognitive state heatmaps.

Cross-stimuli Analysis

First, two sets of three images, reflecting three different levels of perceived visual complexity were created. In the first of the two series, participants were asked to find information about upcoming events that would interest them for a visit to Luxembourg. These images, presented in figure 5, were then analyzed using the cross-stimuli gradient.

When analyzing participants' cognitive load heatmaps, we can see which areas of each interface induced higher levels of cognitive load. More so, when compared to the other images, we can also ascertain which interface induced greater overall cognitive load. Of the three images, the area which inferred greater cognitive load is located at the top left corner of the most complex image (Figure 5 - right-hand side). When comparing this area with the same areas within the other two images (i.e. located in the same spatial location), we can see that the information content has not changed. Yet, the experienced cognitive load in the most complex version of the interface differs from the others. This is in line with the manipulations performed on the

aforementioned image. As mentioned in the Experimental Validation section, changes in color, including the hue, and number of colors, clutter and the quantity of objects were manipulated in order to render the image more complex. The cognitive load heatmaps indicate that these changes influenced participants' perceived visual complexity, therefore experiencing higher cognitive load. Simply put, the number of colors and elements surrounding the area made it more difficult for users to assimilate the information, requiring more cognitive resources.

Looking at Figure 5, again using the cross-stimulus gradient, we can also see that changes in visual complexity also impacted users' behavior. Fixations on the third webpage (right-hand side of Figure 5) seemed to be aggregated in specific locations, in the top half of the interface, when compared to the less complex images where fixations were more widespread (left-hand side). Participants seemed to focus their attention and efforts towards specific areas, when faced with a more complex interface. One of the conclusion we can draw from this visualization is that users' seemed hesitant to explore websites deemed more complex. As stated by Harper et al. 2009 [13], "*Visual complexity seems to be an implicit key into the perceived cognitive load of the page and the interaction that the users think will be required to use the resource*." This statement is in line with our analysis, as UX heatmaps show user's reluctance to explore the page, anticipating the required cognitive load needed to do so.

Figure 9. On the left, negative valence (red) and positive valence (yellow) heatmaps are illustrated. On the right, negative valence (red) and cognitive load (green) heatmaps.

In the context of this task, the comparison of various images using a cross-stimuli gradient allowed us to see the effect of design changes on user experience, both in terms of behavior and experienced cognitive load. For practitioners, this visualization can help make a more informed decision in a prototype comparison or A/B testing context.

Analysis of Emotional and Cognitive States

The Circumplex model of affect describes emotions using the two dimensions of valence and arousal [35]. Valence is used to contrast states of pleasure (e.g., happy) and displeasure (e.g., angry), and arousal to contrast states of low arousal (e.g., calm) and high arousal (e.g., surprise). Using the second set of images, we compared regions of high cognitive load versus regions of negative emotional valence; as well as areas of negative and positive valence.

Figure 9 shows the simplified version of a popular e-commerce website used during Task 2, in which participants were asked to look at the page and select an item they would like to purchase. When looking at the left-hand side of Figure 9, we can see the mapping of a positive emotional valence heatmap (yellow) and a negative emotional valence heatmap (red). Higher intensity emotion areas, or hotspots, are not located on the same pictorial information when comparing the two visualizations. The negative valence heatmap indicates that users experienced displeasure with higher frequency on text and navigation areas, whereas the positive valence heatmap hotspots are located on the human faces at the top of the interface, as well as on the video game (bottom right corner).

The right-hand panel of Figure 9 represents the mapping of a cognitive load heatmap (green) and a negative emotional valence heatmap (red). As we can see, regions of high cognitive load and areas of negative valence overlap. Therefore, in the context of this task, we can observe that high cognitive load, induced by the visual complexity of the page, led users to experience negative emotions. Furthermore, these clustered fixations seem to be located predominantly on text heavy areas.

In this example, comparing regions of high cognitive load to regions of negative emotional valence can help UX practitioners to not only identify the problematic areas of an interface, for instance areas that can adversely affect the user in the completion of a task or that could serve as deterrent to the use of a system, but also contextualize these emotional states by highlighting the graphical elements behind such states.

From a research perspective, such a tool can be used to visualize the effects of high cognitive load on users' emotional state. For instance, the tool's triangulation of UX constructs makes it possible to test the proposed inverted U-shaped relationship between complexity and emotional valence, proposed by Berlyne (1974) [2]. Simply put, interfaces either deemed too simple or too complex, will result in lower affective valence. As shown in Figure 10, results obtained in task 2 are in line with this finding as we can see that areas of high cognitive load overlap areas of negative valence.

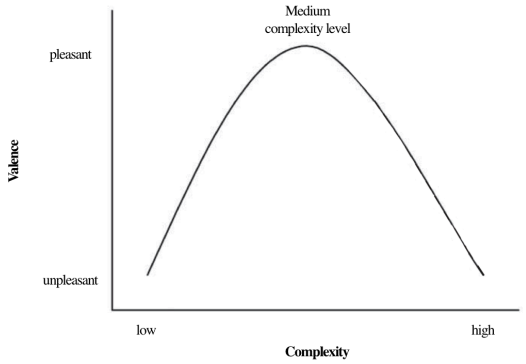

Figure 10. The relationship between visual complexity and affective valence.

LIMITATIONS AND FUTURE WORKS

The proposed tool aims to support the work of experts (i.e., ergonomists, designers, researchers) from the HCI and UX communities. While UX heatmaps can communicate data, they cannot make a diagnosis. To maximize their usefulness, the tool should be integrated to the methods already available to experts (e.g. questionnaires, observations, interviews). Furthermore, as for most eyetracking software, the proposed tool can only be used to analyze interactive interfaces at different points in time, i.e. different stills of the interface, but not interfaces in motion (e.g. flash animations). This represents the main limitation of the tool.

The next step for future work is the inclusion of other cognitive and emotional states, such as emotional UX heatmaps based on physiological signals, as opposed to FaceReader data in the tool's current version. Moreover, we will be conducting interviews with researchers and practitioners in order to evaluate the UX heatmap tool, and to better understand the needs of the UX community for future iterations.

CONCLUSION

The objective of this paper was to present an off-the-shelf, easy to interpret UX evaluation tool which contextualizes users' signals while interacting with a system. Using these signals to infer users' emotional and cognitive states and mapping these states using UX heatmaps on the interface provides researchers and practitioners with a useful tool to contextualize users' reactions. This triangulated approach makes it possible to visually analyze users' various emotional and cognitive states for specific areas of a given interface (e.g., cognitive load combined with emotional valence). Moreover, the cross-stimuli color gradient makes it possible to compare users' gaze patterns and states across multiple interfaces (e.g., A/B testing).

ACKNOWLEDGMENTS

Authors want to thank Brendan Scully for manuscript revision. The authors also want to thank the research assistants who administered the study. This work was supported by the FRQSC (Fonds de Recherche du Québec - Société et Culture).

REFERENCES

1. Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A. and Montanari, R. 2011. Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour* 14, 3: 199-208.
2. Berlyne, D. *Studies in the New Experimental Aesthetics*. Hemisphere Publishing, Washington, DC, 1974.
3. Boucsein, W. *Electrodermal Activity*. Springer, Berlin, 2012.
4. Breslow, L.A., Ratwani, R.M. and Trafton, J.G. 2009. Cognitive Models of the Influence of Color Scale on Data Visualization Tasks. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 51, 3: 321-338.
5. Cacioppo, J.T., Tassinary, L.T. and Berntson, G. Psychophysiological Science : Interdisciplinary Approaches to Classic Questions About the Mind. in Cacioppo, J.T., Tassinary, L.G. and Bernston, G.G. eds. *Handbook of Psychophysiology*, Cambride University Press, New York, 2007, 1-18.
6. Calvo, R.A. and D'Mello, S. 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *Affective Computing, IEEE Transactions on* 1, 1: 18-37.
7. Cernea, D., Weber, C., Ebert, A. and Kerren, A., Emotion scents: a method of representing user emotions on gui widgets. in *IS&T/SPIE Electronic Imaging*, (2013), International Society for Optics and Photonics, 86540F-86540F-86514.
8. Courtemanche, F., Dufresne, A. and LeMoyne, É. Multiresolution Feature Extraction During Psychophysiological Inference: Addressing Signals Asynchronicity. in da Silva, H.P., Holzinger, A., Fairclough, S. and Majoe, D. eds. *Physiological Computing Systems*, Springer Berlin Heidelberg, 2014, 42-55.
9. de Guinea, A.O., Titah, R. and Léger, P.-M. 2014. Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. *Journal of Management Information Systems* 30, 4: 179-210.
10. Dufresne, A., Courtemanche, F., PromTep, S. and Sénécal, S., Physiological Measures, Eye Tracking and Task Analysis to Track User Reactions in User Generated Content. in *Measuring Behavior*, (Eindhoven, The Netherlands, 2010), 218-222.
11. Ganglbauer, E., Schrammel, J., Deutsch, S. and Tscheligi, M. Applying Psychophysiological Methods for Measuring User Experience: Possibilities, Challenges and Feasibility *User Experience Evaluation Methods in Product Development (UXEM'09) workshop at the 12th IFIP TC13 Conference on Human-Computer Interaction*, 2009.
12. Haapalainen, E., Kim, S., Forlizzi, J.F. and Dey, A.K. Psycho-physiological measures for assessing cognitive load *Proceedings of the 12th ACM international conference on Ubiquitous computing*, ACM, Copenhagen, Denmark, 2010, 301-310.
13. Harper, S., Michailidou, E. and Stevens, R. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception (TAP)* 6, 2: 10.
14. Hassenzahl, M., Diefenbach, S. and Göritz, A. 2010. Needs, affect, and interactive products–Facets of user experience. *Interacting with computers* 22, 5: 353-362.
15. Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R. and Jarodzka, H. *Eye Tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.
16. Huisman, G., Hout, M.v., Dijk, E.v., Geest, T.v.d. and Heylen, D. LEMtool: measuring emotions in visual interfaces *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, Paris, France, 2013, 351-360.
17. Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J.P., Karvonen, H., Keskinen, T., Koskinen, H., Lu, Y. and Saariluoma, P. 2015. Defining user experience goals to guide the design of industrial systems. *Behaviour & Information Technology*, ahead-of-print: 1-16.
18. King, M.F. and Bruner, G.C. 2000. Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing* 17, 2: 79-103.
19. Kivikangas, M., Nacke, L. and Ravaja, N. 2011. Developing a triangulation system for digital game events, observational video, and psychophysiological data to study emotional responses to a virtual character. *Entertainment Computing* 2, 1: 11-16.
20. Kreibig, S.D. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3: 394-421.
21. Laeng, B., Sirois, S. and Gredebäck, G. 2012. Pupillometry. *Perspectives on Psychological Science* 7, 1: 18-27.
22. Nielsen, J. and Pernice, K. *Eyetracking Web Usability*. New Riders, Berkeley, California, 2012.
23. Oliva, A., Mack, M.L., Shrestha, M. and Peeper, A., Identifying the perceptual dimensions of visual complexity of scenes. in *26th Annual Cognitive Science Society*, (Austin, TX, 2004).
24. Olkin, I. and Finn, J.D. 1995. Correlations redux. *Psychological Bulletin* 118, 1: 155.
25. Pantic, M., Nijholt, A., Pentland, A. and Huanag, T.S. 2008. Human-Centred Intelligent Human-Computer Interaction (HCI2): How far are we from attaining it?

International Journal of Autonomous and Adaptive Communications Systems 1, 2: 168-187.
26. Rayner, K. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin.* 124, 3: 372-422.
27. Robinson, M.D. and Clore, G.L. 2002. Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology* 83, 1: 198-215.
28. Roto, V., Vermeeren, A.P.O.S., Väänänen-Vainio-Mattila, K., Law, E. and Obrist, M., Course notes: User Experience Evaluation Methods - Which Method to Choose? in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Paris, France, 2013), ACM.
29. Shi, Y., Ruiz, N., Taib, R., Choi, E. and Chen, F. Galvanic skin response (GSR) as an index of cognitive load *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, ACM, San Jose, CA, USA, 2007, 2651-2656.
30. Sweller, J. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2: 257-285.
31. van den Broek, E., van der Zwaag, M.D., Healey, J.A., Janssen, J.H. and Westerink, J.H.D.M. Prerequisites for Affective Signal Processing (ASP) - Part IV *1st International Workshop on Bio-inspired Human-Machine Interfaces and Healthcare Applications - B-Interface 2010*, Valencia, Spain, 2010, 59-66.
32. Vermeeren, A.P.O.S., Law, E.L.-C., Roto, V., Obrist, M., Hoonhout, J., V, K., #228 and nen-Vainio-Mattila. User experience evaluation methods: current state and development needs *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ACM, Reykjavik, Iceland, 2010, 521-530.
33. Wang, Q., Yang, S., Liu, M., Cao, Z. and Ma, Q. 2014. An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems* 62: 1-10.
34. Ward, R.D. and Marsden, P.H. 2003. Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies* 59, 1-2: 199-212.
35. Wooding, D.S. Fixation maps: quantifying eye-movement traces *Proceedings of the 2002 symposium on Eye tracking research & applications*, ACM, New Orleans, Louisiana, 2002, 31-36.
36. Zaman, B. and Shrimpton-Smith, T. The FaceReader: measuring instant fun of use *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, ACM, Oslo, Norway, 2006, 457-460.
37. Zhihong, Z., Pantic, M., Roisman, G.I. and Huang, T.S. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1: 39-58.