http://dx.doi.org/10.14236/ewic/HCI2016.25

Assessing User Interface Aesthetics based on the Inter-subjectivity of Judgment

Mathieu Zen
Université catholique de Louvain
Louvain School of Management, LiLab
Place des Doyens, 1 - 1348 Louvain-la-Neuve, Belgium
mathieu.zen@uclouvain.be

Jean Vanderdonckt
Université catholique de Louvain
Louvain School of Management, LiLab
Place des Doyens, 1 - 1348 Louvain-la-Neuve, Belgium
jean.vanderdonckt@uclouvain.be

"How to assess user interface aesthetics?" remains a question faced by many user interface researchers and designers during the user interface development life cycle since aesthetics positively influence usability, user experience, pleasureability, and trust. Visual techniques borrowed from visual design suggest that the graphical user interface layout could be assessed by aesthetic metrics such as balance, symmetry, proportion, regularity, and simplicity, to name a few. Whereas different formulas exist for computing each aesthetic metric and different interpretations to sum up their results, no consensus exists today on how to consistently evaluate these metrics in a way that is aligned with human judgement, which is intrinsically subjective. In order to address the challenging alignment of human subjectivity with machine objectivity, this paper reports on an experiment comparing the results issued from the inter-subjectivity of judgment of fifteen participants evaluating four main aesthetic metrics on a sample of ten graphical user interfaces and the values of these metrics calculated semi-automatically by a web-based application. The experiment suggests that some metrics, e.g. symmetry, proportion, simplicity, as computed from the formula are actually positively correlated with human judgment, while some other metrics, such as balance, are surprisingly not correlated with the formula computed, thus indicating that another formula or another interpretation should be determined. Therefore, a new formula for computing balance is defined that decomposes balance into horizontal and vertical balances which re-establish a correlation. This paper then provides some new insights on how to rely on these aesthetic metrics and other related metrics, whether they are interpreted manually or computed automatically.

User interface aesthetics, metric-based evaluation, theory of measurement, visual design, visual principles

1. INTRODUCTION

"Not only in work definitely undertaken from artistic impulse, but in all the products of his industry, in his choice of locality, in his dwelling, his clothes and his implements, man shows that he is affected by appearance, by something that causes him pleasure over and above the immediate utility of object."

This quote from Felix Clay (1908) illustrates that design matters when it comes to make a choice and that appearance represents a competitive advantage in industry. Several organisations have integrated this principle throughout their product development life cycle. For instance, they do not only improve packaging or count on originality for reaching a high-level of user satisfaction, because beauty and design make the sale, aesthetics became their primary concern.

Aesthetics are often defined as a subjective matter related to the concept of "beauty", "visual design", "appealing", which are abstract concepts mainly studied in disciplines such as philosophy, psychology, social sciences, and arts. In general, aesthetics are considered as "the immediate pleasurable experience that is desired toward an object and not mediated by intervening reasoning" (Moshagen and Thielsch (2010)). In Human-Computer Interaction (HCI), user interface aesthetics or visual aesthetics are mainly referred to as the beauty or the pleasing appearance of interacting a user interface of an interactive system (Tractinsky et al. (2000); Robins and Holmes (2008))

On the one hand, evaluation is by nature an objective process resulting into quantifiable results, while, on the other hand, the human judgment, especially when it comes to evaluating aesthetics, remains by nature a subjective process resulting into qualitative results. To address the challenge posed by this duality, one option consists in considering the concept of inter-subjectivity stated by Kant (1987) as the idea that all human are thinking subjects able to take into consideration others' thought in their own judgment. As a result, judgments rendered in a given context are not isolated

and are likely to be reproduced in another place at another time, thus allowing to gather the subjective judgments for extracting a more objective response (Figure 1). As opposed to solipsistic individual user experience in which each user has her own particular experience that is not necessarily shared with others, inter-subjective collective user experience is expected to reach some agreement between users on a given set of meanings or a definition of the situation. With respect to aesthetics, we hereby define *inter-subjectivity of judgment* as the process of collecting the subjectivity of a set of participants evaluating the aesthetics of a user interface in order to produce a quantitative assessment that is agreed upon by participants. Inter-subjectivity, respectively no inter-subjectivity, will be reached when a agreement, respectively a disagreement, among participants will emerge from assessing an aesthetic metric of a user interface.

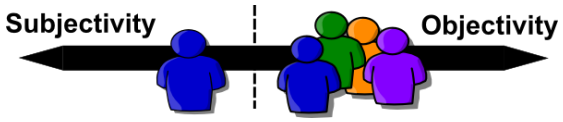


Figure 1: *Inter-subjectivity of judgement*

The goal is not to discuss if the aesthetic judgement is subjective or not - it certainly is for any person in any defined context - but rather to find criteria of this judgement that are common to a large number of individuals. Based on the concept of inter-subjectivity, we assume that any individual originating a user interface is able to “feel” those common criteria and take them into account in her creative artwork. This claim could be expressed as: aesthetics are subjective from a personal point of view, but the aggregation of a set of judgments enables determining a certain degree of objectivity or at least consider common elements of inter-subjectivity. By collecting different opinions on the aesthetics of an object, a work or any exposure, it is possible to get an answer that is collectively objective. This is a premise to consider when assessing aesthetics.

2. THEORETICAL BACKGROUND

2.1. Graphical User Interface Aesthetics

User Interface aesthetics are considered as a key factor having a direct relation with the perceived usability of the user interface (Ben-Bassat et al. (2006); Tuch et al. (2012)), the effective usability (Hamborg et al. (2014)), its testing (Sonderegger and Sauer (2010); Reinecke et al. (2013)), on the subjective user satisfaction (Tractinsky et al. (2000)), on the visual search performance (Salimun et al. (2010a)), on the first impression flashed by the user interface (Lindgaard et al. (2006)) or its preference (Schenkman and Jönsson (2000)). End users could put aesthetics ahead of usability (Papachristos and Avouris (2011)), even when they are formally asked to describe usability only (Kurosu and Kashimura (1995)). Moreover, their first impression about a UI takes less than 50 milliseconds to be formed (Lindgaard et al. (2006)) based namely on aesthetics. Norman (2005) differentiates three levels of design judgment: visceral, behavioral and reflective. The visceral experience occurs when the design appears to the user who creates an immediate opinion about it, i.e. she likes it or not. The behavioral experience is forged during the use of a design, the most cognitive experience. Afterwards comes the reflective experience when the user has made his own opinion about the design - taking into account not only the satisfaction but also some emotional attachment induced by the design - and makes his choice to continue using it. It explains why people can confine themselves using non effective objects because they have created with it an emotional link. These arguments may also be used to another universe than only one involving design and packaging since it had already been demonstrated with the Dion’s "halo effect" (Dion et al. (1972)) that people considered physically attractive had more chances to be perceived competent and successful. Higher aesthetic treatments produce higher judgments of credibility in users’ mind (Robins and Holmes (2008)).

2.2. Aesthetic Metrics

A typical approach followed to characterise aesthetics, which is intrinsically subjective, consists in trying to make the analysis as objective as possible by replacing it by a series of metrics that, taken together, are supposed to cover the original concept. This approach has been materialised into a set of low-level metrics (e.g. amount of elements, size of elements, amount of colors, such as in Sherlock (Mahajan and Shneiderman (1997)), one high-level metric at a time (such as the Layout Complexity (Comber and Maltby (1997)) or the Layout Appropriateness (Sears (1993)) or a set of high-level metrics (such as metrics borrowed from visual design). When aesthetics are attempted to be characterised by a set of metrics, some approaches are aimed at minimizing the amount of metrics required to grasp aesthetics as a whole (e.g. minimizing the amount of descriptors) while some other tend to open the range of possible ways to analyze the same user interface against a set of metrics (Purchase et al. (2011)).

Vanderdonckt and Gillo (1994) introduced a set of 30 visual techniques for designing graphical user interfaces that fall into five categories: *physical techniques* (e.g. balance, symmetry, alignment), *composition techniques* (e.g. simplicity, economy, understatement), *association and dissociation techniques* (e.g. unity, repartition, grouping), *ordering techniques* (e.g. consistency, predictability, sequentiality), and *photographic techniques* (e.g. sharpness, roundness, opacity). For each visual technique, two extremes are located at the opposite edges of a continuum: balance is located on the side

of harmony and instability is located on the side of contrast. Similarly, symmetry is located on the side of harmony while asymmetry is located on the side of contrast. This does not mean that an asymmetric UI is necessarily considered unusable: it means that the asymmetry will contribute to contrast as opposed to harmony, thus leading to consequences on the UI which may be desirable (e.g. in order to increase attractiveness for a game) or prejudicial (e.g. risk on decreasing the productivity of an information system). The main shortcoming of these definitions is that their calculation and their interpretation are left to the responsibility of the person using them, which may vary from one person to another depending on their level of expertise.

2.3. Computable Aesthetic Metrics

In order to address this shortcoming, Ngo et al. (2000) were the first to state a mathematical formula for calculating these visual principles by computable aesthetic metrics based on a particular interpretation of these visual principles or other properties. For any particular aesthetic metric, a multitude of heterogeneous, potentially inconsistent, formula exists along with their own interpretation. For instance, symmetry and balance are famous for their extensive handling in the literature through a large set of formulas, interpretations, and algorithms for calculating them. This paper will focus on Ngo et al. (2000, 2003)'s formula since they have reached a reasonable level of acceptance and recognition in the scientific community and they have been subject to several empirical studies. Two representative examples are: (i) Salimun et al. (2010b) computed six aesthetic metrics, i.e. cohesion, economy, regularity, sequence, symmetry, and unity and analyzed their values based on forced-choice paired comparisons to conclude that some metrics, such as symmetry and cohesion, are more influential than others; (ii) Mõttus et al. (2013) manually extracted UI elements from six web pages, manually computed some Ngo's formulas, and compared these scores with appreciations of participants to conclude that correlations between them were weak.

2.4. Software for Computing Aesthetic Metrics

Since a wide variety of aesthetic metrics exist with their own formula and interpretations, it makes sense to develop a software that (semi-) automatically computes the values of these metrics (Zain et al. (2011)) so as to release the evaluator from computing them manually, which is a tedious task, and to avoid that this process would be error prone. Another shortcoming that needs to be addressed by the software would be the speed of deployment since existing tools evaluating UI aesthetics still require to be downloaded and sometimes compiled and installed before being ready to use. In addition, the quantitative value of computed metrics (*summative evaluation*) could be augmented by a qualitative approach in which designers are provided with suggestions on how to improve the UI (*formative evaluation*) so as to maximise the value of these metrics. Before doing so, it is necessary to align the metrics as they could be computed by software to the ones that people have in mind, subjectively or objectively. For this purpose, two approaches are typically considered:

1. **Infer computable metrics from human judgement**: if people have in mind a certain understanding of the aesthetic metrics, even if subjective, they should be able to rank UIs by decreasing order of these metrics and assign a potential value to these metrics. A software may then be trained by machine learning to produce a value for each metric that is as close as possible to the human judgement. This approach has two shortcomings: a machine learning process is expensive and sometimes hard to conduct because of unavailability of people and the resulting score produced by the system cannot be neither explained nor enter in a more formative process.
2. **Derive human judgement from computed metrics**: if people see a particular formula computed for any aesthetic metric, they may have their own interpretation of the formula, thus leading to a human judgement that is possible to capture Zain et al. (2011). This approach does not suffer from the two aforementioned shortcomings: no machine learning process is required to train the software until an acceptable data set is reached and the formula keeps its explanatory power, especially for formative evaluation. Therefore, determining to what extent these formula, as computed by a software, are aligned with the corresponding human judgement is important, which is the purpose of this study.

Judgement is considered subjective and personal. Hence, validating metrics in order to give credit to the evaluation process itself is crucial. The purpose of this study is to experimentally compare the machine versus the human aesthetic evaluation in order to see how reliable the software ratings are.

3. COMPARATIVE STUDY

In order to give credit to the automatic computing of aesthetic metrics by a software, the aesthetic metrics, as they are computed, need to be aligned with the human judgement. This section reports on the results of an experiment conducted towards this goal.

3.1. Goal

In order to identify any relation between the scores computed for aesthetic metrics and the reviews obtained from human judgment, the following methodology has been applied:

- Select a subset of aesthetic metrics: instead of computing a potentially large set of aesthetic metrics, like the 30 visual techniques in Vanderdonckt and Gillo (1994), a subset of significant and representative metrics should be selected.
- Consider an existing set of real-worlds UIs exhibiting heterogeneous aspects: instead of generating random UIs based on the same set of UI elements, it was preferred to rely on real-world UIs coming from frequently used web sites.
- Compute formula associated to the selected metrics for each UI.
- Conduct a comparative study with users to collect data about their perceptions of the techniques.
- Compute descriptive statistics and identify correlations based on inter-subjectivity of judgement.

3.1.1. Comparative method

Unlike other research experiment conducted in the past (Mõttus et al. (2013)), we chose not to present to the participants one UI at a time and asking them to evaluate its aesthetic metrics separately (Figure 2a) in order to minimise repetitive tasks that would lead to participant's fatigue. We noticed that this method was making each UI being evaluated independent from each other. Therefore, we chose to conduct an experiment with a comparison method. This approach benefits from several potential advantages over previously conducted methods: participants do not need to remember the scores they previously assigned to other UIs, thus avoiding inconsistent and conflicting results; the method simplifies the evaluation process dramatically by avoiding to ask each participant to evaluate each aesthetic metric for each UI considered in the data set. Instead of giving a score for one metric on a 5-point Likert scale or quantitatively for one UI at a time, the participants were instructed to pick from a pair of two candidate UIs which one was representing the best the aesthetic metric, i.e. that would have obtained a higher score for the measure (Figure 2b).

3.2. Description

3.2.1. Participants

Participants were recruited from a mailing list maintained at Université catholique de Louvain. No compensation was offered to volunteers. The experiment took place in a controlled environment where respondents (N=15) replied to an electronic survey in the presence of interviewers. Before starting the experiment, the participants were asked to provide some personal information for statistics:

- age;
- gender;
- background;
- prior experience with aesthetics;
- prior experience with user interface evaluation.

Each participant daily uses a computer or an electronic device with a graphical user interface. There were six females and nine males, with an average of 28.4 years old (σ=10.08, $\bar{x}$=24), from which only few of them (N=3) reported to have prior experience with UI evaluation. Six participants had a diploma in a scientific degree whereas nine participants had a degree in social sciences. All of them were comfortable with the use of a computer as it is generally one of their daily work tool. Web browsing is a regular activity for the participants.

3.2.2. Aesthetic metrics

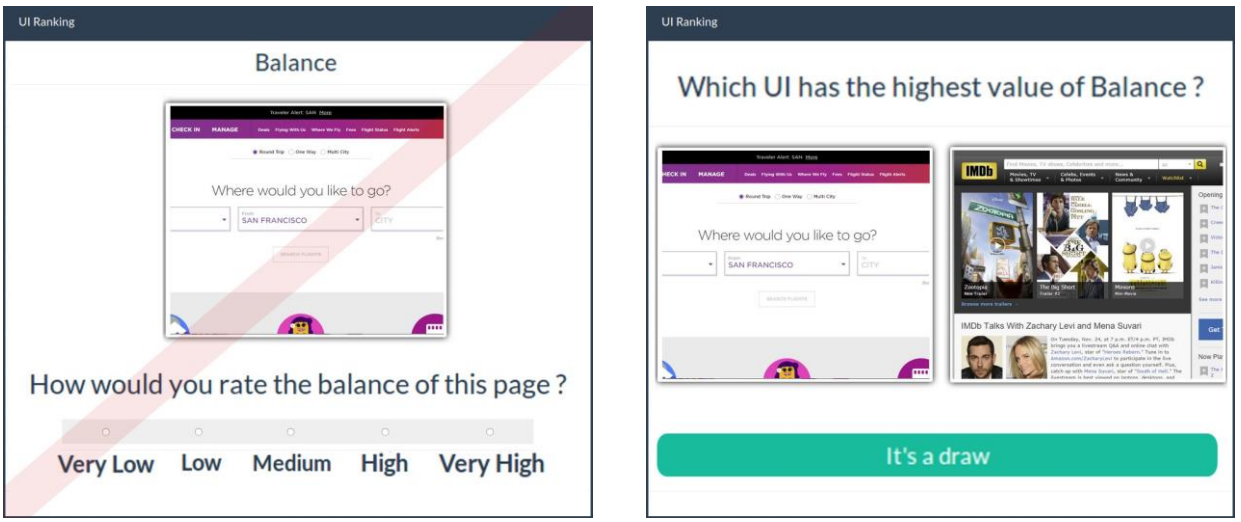
After this initial phase, the aesthetic metrics, as formalised by Ngo et al. (2000), were briefly explained and described to the participants in the same manner:

- *Balance*: is a search for equilibrium along a vertical or horizontal axis in the UI layout.
- *Symmetry*: consists of duplicating the visual image of UI consumption along a horizontal and/or vertical axis or any other axis they would identify.
- *Proportion*: an aesthetically appealing ratio between the dimensions of UI elements belonging to selected UIs. The ratio is calculated by dividing the height of any element by its length. We did not emphasise any particular reference ratio, like the golden proportion (1.6180).
- *Simplicity*: is directness and singleness of UI layout, free from secondary complication.

In order to keep the simplicity of the experiment and to respect imposed time constraints, we focused on aesthetic metrics that were considered the most obvious for end users and the most frequently researched in the literature or mentioned in textbooks on visual design. Results from Altaboli and Lin (2011) showed that balance, unity, and sequence have significant effects on the perceived interface aesthetics and reveal a significant interaction among them. The aforementioned techniques were immediately understandable by participants since none of them asked for further explanation.

3.2.3. Training pre-test

In order to ensure some internal validity and let the participants become familiar with the experimental design, a preliminary test took place with the same policy, rules, and layout than the effective one, but with only four UIs. At the end of the pre-test, participants were able to see the resulting ranking of UIs for each metric they evaluated and compare their own reviews with the aggregated reviews of the participants, as in a focus group (Figure 3). The primary goal of the pre-test was therefore to increase the degree of expertise of the participants and collect accurate responses in the main phase of the experiment. UIs selected for the pre-test were different from those of the final experiment.



(a) *Score on a 5-point Likert scale scale* ***(b)*** *A-B testing*

Figure 2: *Experimental methods for collecting participants' reviews*

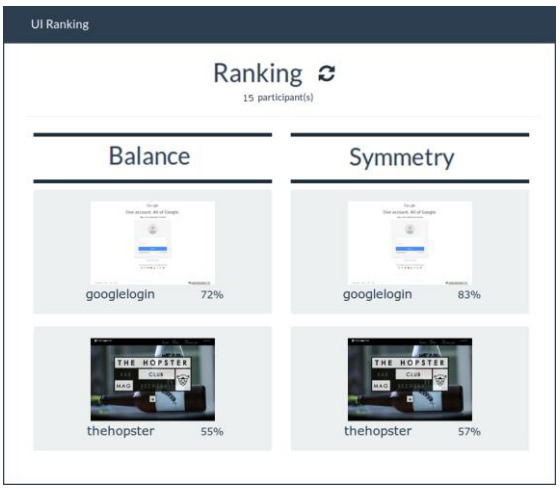


Figure 3: *Pre-test final ranking as conveyed to the participants*

3.2.4. Selected UIs

Ten different web sites were selected (Figure 4) first for their general interest (thus subsuming the fact that no semantic knowledge of the domain is required) and frequency of use, then divided into four categories (i.e. news, on-line shop, education, and booking). These UIs have been selected based on Nielsen and Molich (1990) heuristics about aesthetics and minimalistic design: there are both "beautiful" web sites and "ugly" web sites, where ugly stands for non-compliant with the aesthetic metrics and Nielsen heuristics. In particular, we hypothesised that the web sites are ranked according to their aesthetics score as follows:

- Low: SNCB (railways web site), Yahoo Answers, IMDB cinematographic web site;
- Medium: Erasmus Student Network, Understood, Empire;
- High: Air B&B, Element, Virgin America, BBC News.

All web sites were presented to the participants in the same language (English) since some of them have different mother tongues.

3.2.5. Procedure

We developed and installed an A/B testing survey application (Figure 2b) where two UI screenshots were randomly selected from the pool and were presented individually to the user so as to select by clicking on it which one is representing positively the most the considered aesthetic metric, i.e. the user has to select the one that she thinks to have the higher value for the metric. For every positive selection, one point is given to the global score of the UI. If the participant is undecided, he has the possibility to click on the "draw" button, no point is assigned and two new UIs are further presented. In order to prevent responses biases, the UIs are also displayed in a random order and each pair of UI is verified as being unique to avoid any duplicate. The comparisons are done for each of the four aesthetic metrics. Before each suite of pair comparisons, a textual description with illustration of the aesthetic metric is delivered.

3.2.6. Summary

In a nutshell, the survey requested each respondent to compare the 10 selected UIs two by two in the light of four selected aesthetic metrics. For each metric, forty-five unique comparisons were generated, which could be achieved in more or less five minutes, that is a total of 20 minutes per participant. These data assets allowed us to infer implicit inter-subjective responses since the participants were not asked to produce any explicit group reply during the experiment. No time constraint were

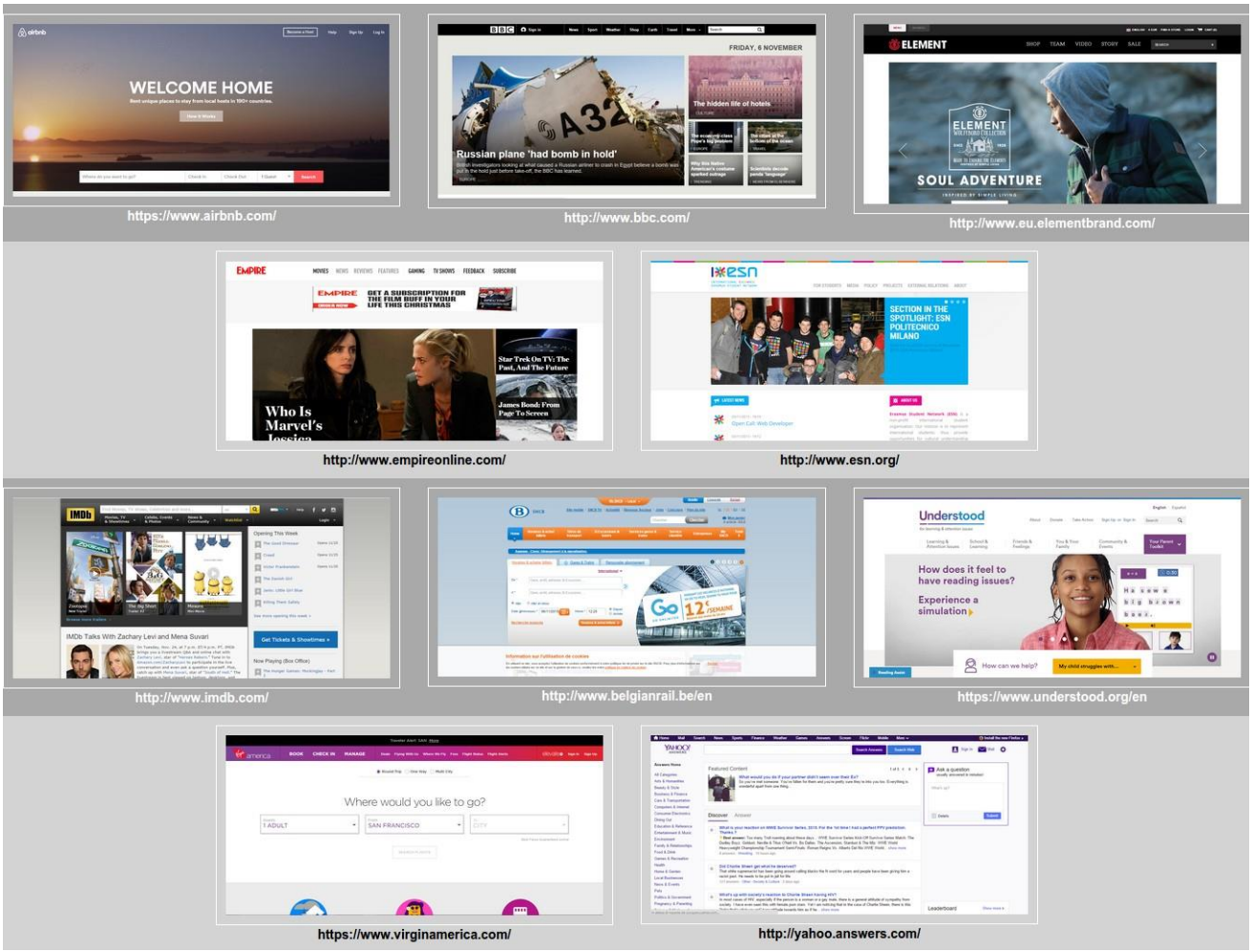


Figure 4: *Set of selected user interfaces*

imposed to participants. The amount of data generated by this experiment is described by the following equation:

15 *users*
× 4 *visual techniques*
× 45 *comparisons*

= 2700 *samples*

3.2.7. QUESTIM

For computing the aesthetic metrics of this experiment, we developed and used QUESTIM (Quality Estimator using Metrics), a Web-based evaluator software enabling semi-automatic computation of the aesthetic metrics as described. QUESTIM is threefold: (i) a web service written in Java and then compiled into JavaScript with Google Web Toolkit (GWT) providing so-called *Aesthetic Evaluation as-a-Service* (AestaaS), which could therefore be reused by another application for the same purpose; (ii) an on-line application based on this web service freely accessible at http://questimapp.appspot.com with which it is possible to copy/paste any URL of a web page to be analysed - therefore, no need to redraw anything; and (iii) an API providing the possibility to post a structured description of the positions and dimensions of layout's regions of interests and immediately receive a JSON response with the computed metrics, which is particular useful for assessing the aesthetics of any graphical user interface (not just web pages), any sketched user interfaces (e.g. produced with wireframe tools such as Balsamiq), hand-drawn user interfaces and other types of UI prototypes (e.g. pictures, excerpts from a video) as opposed to real screenshots. The aesthetic metrics were implemented based on Ngo's formula and summarised as follows:

- For computing *balance*, we mainly implemented the formula given by Ngo et al. (2000) but chose to focus only on the elements dimension and position properties, thus leaving aside objects colors and complexity. The impact of this choice is further elaborated in the discussion section. We also inferred some part of the algorithm that were left vague in Ngo's formula, specifically the computation of the distance between UI objects

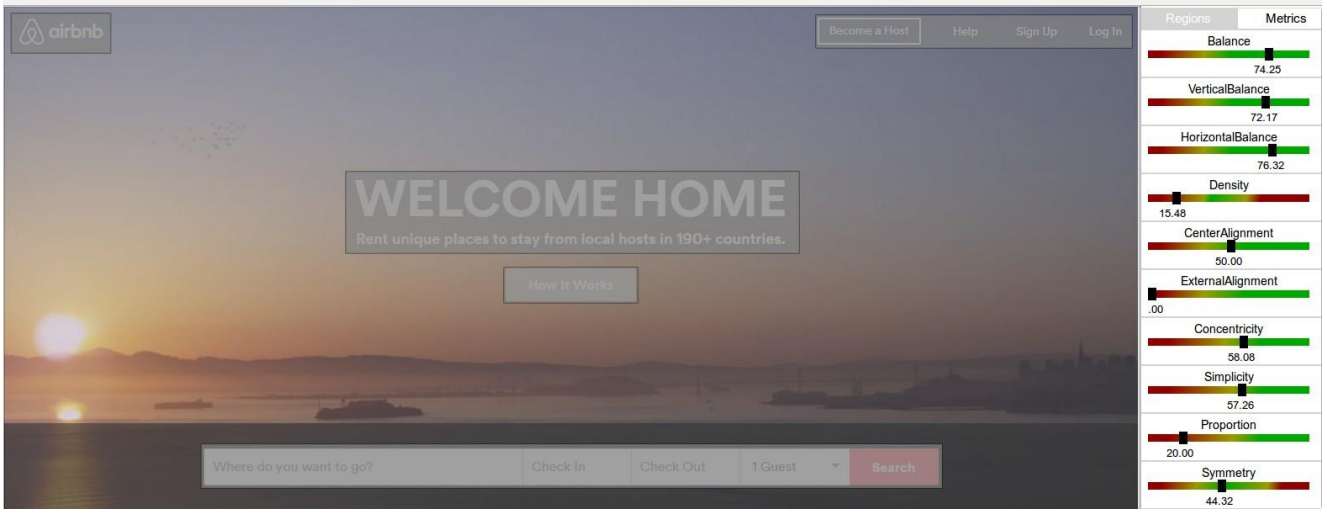


Figure 5: *QUESTIM computation process*

and the frame as well as the strategy for dividing an object over-spanning two sections which is simply splitting the latter in two distinct sub-objects with their own area and center.

- The metric for *simplicity* is close to the one proposed by Ngo et al. (2000) while being slightly different as it takes into account an alignment property, a degree of similarities between the objects and a measure of the layout density. In short, the metric considers a UI to be simple if its objects are both aligned vertically and/or horizontally, relatively similar and their arrangement is sparse on the screen, possibly enriched by other properties like horizontal/vertical centering, equilibrium or uniformity.
- *Symmetry* is a straightforward application of what is provided by Ngo et al. (2000).
- *Proportion* measurement method only focuses on the degree of similarity between objects dimensions ratio, e.g. UI with all objects being squares should give a high score for this metric.

In order to collect a report including the scores given by QUESTIM for the aforementioned metrics, a semi-automatic computation consists in two steps (Fig. 5):

1. The UI is loaded on the screen and the experimenter defines manually the layout regions of interests. Note that an edge detection filter could be applied for semi-automated region detection and segmentation, but we preferred to do them manually to avoid any irrelevant region.
2. Based on the position (x and y) and dimensions (width and height) of each object derived from the regions, QUESTIM computes the values of balance, simplicity, proportion, and symmetry. The values are then represented as a bar on a slider with colors ranging from red (lowest score) to green (highest score) so as to convey not only the value of the metric (summative evaluation), but also its interpretation and its positioning (formative evaluation). Note that the color gradients vary depending on the metric, based on accumulated computations related to usability.

3.3. Results

3.3.1. Inter-judge agreement

In order to assess the internal validity, we chose first to evaluate degree of concordance in participants responses using Randolph's Kappa coefficient Randolph (2005). For this purpose, the following parameters were considered: the forty-five unique comparisons (i.e. the cases in the sense of Randolph), the fifteen raters and three categories: A>B, B>A and A=B (where A and B are the respective compared UIs). The input were therefore the number of raters who agreed that a certain case belonged to a certain category. For each case, the sum of the categories had to reach to an amount of fifteen (the total number of raters). Table 1 reproduces the results of this calculation: where as the overall agreement remains above 0.5 with 0.64 for aesthetics, fixed-marginal and free-marginal Kappa's values are varying: between 0.49 for the highest and 0.25 for the lowest for the free-marginal kappa which means that the inter-judge agreement is not above the threshold of 0.7, but that there exists some agreement between judges not random (>0). Randoplh's kappa coefficient is estimated more restrictive than Kendall's Kappa coefficient, which will be considered later on.

Table 1: *Randolph's Kappa coefficients*

	Aeshetics	Balance	Simplicity	Symmetry	Proportion
Overall agreement %	0.649523	0.505396	0.66074	0.582222	0.507724
Fixed-marginal kappa	0.263656	0.110834	0.353084	0.258599	0.134831
Free-marginal kappa	0.474285	0.258094	0.49111	0.373333	0.261586

3.3.2. Aesthetics Score

As a second step, we analyzed the data gathered exclusively about the global aesthetics of the pages. The participants were asked to judge only how pleasant were the screens for them. Using the same methodology discussed supra, namely a paired comparison model, participants were asked to choose betweenUIs which one they found the most visually pleasant. Therefore, we were able to draw a symmetric matrix of results (Table 2) associating for each pair of UI_A and UI_B a score between +15 and -15 (15 being the number of participants) where a score of +15 would mean that UI_A is preferred to UI_B by all respondents in terms of aesthetics. Whereas a score of -15 would mean that UI_B is preferred to UI_A. We also used the raw data obtained through the experiment in order to compute a latent score of aesthetics for each UI. For this, we used the Bradley-Terry-Luce (BTL) model (Courcoux and Semenou (1997)) in order to establish a ranking attached with a score for each UI giving an order of magnitude to the same ranking. Basically, we organised the raw data in order to create a preference vector for each UI containing the matches with the nine other UIs (+1 if preferred, -1 if unpreferred) and then applied the BTL method summing the probabilities for one UI to be preferred to all others. This sum of probabilities can be seen as a latent score for the UI. Table 3 reproduces the ranking of each UI with its computed score.

Table 2: *Aesthetics values (participants)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	11	6.0	13	13	13	15	13	9.0	13
bbc	-11	0	-6.0	6.0	8.0	12	11	3.0	2.0	15
elem	-6.0	6.0	0	8.0	14	15	11	10	6.0	15
empi	-13	-6.0	-8.0	0	7.0	11	11	1.0	1.0	15
esn	-13	-8.0	-14	-7.0	0	0.0	7.0	-7.0	-11	11
imdb	-13	-12	-15	-11	0.0	0	-2.0	-9.0	-12	10
sncb	-15	-11	-11	-11	-7.0	2.0	0	-6.0	-8.0	8.0
unde	-13	-3.0	-10	-1.0	7.0	9.0	6.0	0	-3.0	13
virg	-9.0	-2.0	-6.0	-1.0	11	12	8.0	3.0	0	15
yaho	-13	-15	-15	-15	-11	-10	-8.0	-13	-15	0

Table 3: *UI aesthetics sorted by magnitude*

UI	Score	Experimenters classification
airbnb	.484	HIGH
element	.221	HIGH
bbc	.085	HIGH
virginAmerica	.072	HIGH
empire	.058	MEDIUM
understood	.045	MEDIUM
esn	.014	MEDIUM
sncb	.009	LOW
imdb	.009	LOW
yahoo	.002	LOW

3.3.3. Simplicity score and comparison with the software

After normalizing the values from table 2, a new table is obtained with scores between -1 and +1. This normalised table for the value of aesthetics would not be useful as the tool does not give a global value of aesthetics but rather a specific value for its visual components, namely for the aesthetic metrics. We chose therefore to present from now the values for the simplicity, the principle is the same as above: one counting matrix and one normalised matrix (Table 4).

Table 4: *Normalised simplicity values (participants)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	0.60	0.53	1.0	1.0	1.0	1.0	1.0	0.33	0.87
bbc	-0.60	0	-0.33	0.33	0.60	0.87	0.60	0.47	-0.47	0.60
elem	-0.53	0.33	0	0.47	0.60	0.87	1.0	0.80	-0.067	0.87
empi	-1.0	-0.33	-0.47	0	0.27	0.87	0.60	0.67	-0.60	0.53
esn	-1.0	-0.60	-0.60	-0.27	0	0.47	0.40	0.0	-1.0	0.67
imdb	-1.0	-0.87	-0.87	-0.87	-0.47	0	-0.33	-0.73	-1.0	-0.27
sncb	-1.0	-0.60	-1.0	-0.60	-0.40	0.33	0	-0.27	-1.0	-0.067
unde	-1.0	-0.47	-0.80	-0.67	0.0	0.73	0.27	0	-1.0	0.40
virg	-0.33	0.47	0.067	0.60	1.0	1.0	1.0	1.0	0	1.0
yaho	-0.87	-0.60	-0.87	-0.53	-0.67	0.27	0.067	-0.40	-1.0	0

In order to compare the humans reviews with the scores computed by QUESTIM, we had first to find a common ground for analysis, i.e. a similar data structure. Hence, we used the software considered as a new participant and made it pass the same experiment than the one done by the human participants. We used the difference between the scores given for the metric for each UI as a magnitude scale and normalised it in order to obtain values in a range between -1 and 1, and not exactly -1 or 1. Table 5 reproduces the score given by the software for the metric of simplicity for each pair of UIs.

Table 5: *Normalised simplicity values (software)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	0.65	1.0	0.62	0.41	0.85	0.65	0.44	0.029	0.74
bbc	-0.65	0	0.35	-0.029	-0.24	0.21	0.0	-0.21	-0.62	0.088
elem	-1.0	-0.35	0	-0.38	-0.59	-0.15	-0.35	-0.56	-0.97	-0.26
empi	-0.62	0.029	0.38	0	-0.21	0.24	0.029	-0.18	-0.59	0.12
esn	-0.41	0.24	0.59	0.21	0	0.44	0.24	0.029	-0.38	0.32
imdb	-0.85	-0.21	0.15	-0.24	-0.44	0	-0.21	-0.41	-0.82	-0.12
sncb	-0.65	0.0	0.35	-0.029	-0.24	0.21	0	-0.21	-0.62	0.088
unde	-0.44	0.21	0.56	0.18	-0.029	0.41	0.21	0	-0.41	0.29
virg	-0.029	0.62	0.97	0.59	0.38	0.82	0.62	0.41	0	0.71
yaho	-0.74	-0.088	0.26	-0.12	-0.32	0.12	-0.088	-0.29	-0.71	0

For the sake of clarity and to provide another view of the matrix correlations (comparing participants values and software values), we chose to present this table with heat maps where each cell follows this coding scheme:

- If the sign of a paired comparison value given by QUESTIM is equal to the sign of the same value given by the participants for the same pair, e.g. positive and positive - or - negative and negative, the cell will be colored in green. Otherwise, it will be colored in red.
- The intensity of the cell color is determined by the proximity of the compared values, e.g. if the sign is the same and the absolute values are really close, the cell will be tinted in green with more saturation, while keeping the same hue.

Based on this comparison, we notice that both matrices have a rather high correlation rate. Indeed, except for the case of element and BBC, other UIs simplicity measures are correctly predicted by the software and represent quite well the participants reviews when they were asked

Table 6: *Normalised balance values (participants)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	0.47	0.67	0.53	0.27	0.47	0.20	0.20	0.0	0.33
bbc	-0.47	0	-0.13	0.33	0.40	0.47	0.27 0	.067	-0.13	0.33
elem	-0.67	0.13	0	0.27	0.067	0.60	0.13	0.0	-0.27	0.47
empi	-0.53	-0.33	-0.27	0	0.40	0.40	-0.27	-0.67	-0.53	0.47
esn	-0.27	-0.40	-0.067	-0.40	0	0.40	-0.20	-0.47	-0.87	0.47
imdb	-0.47	-0.47	-0.60	-0.40	-0.40	0	-0.60	-0.40	-0.87	0.0
sncb	-0.20	-0.27	-0.13	0.27	0.20	0.60	0	0.0	-0.27	0.73
unde	-0.20	-0.067	0.0	0.67	0.47	0.40	0.0	0	-0.33	0.60
virg	0.0	0.13	0.27	0.53	0.87	0.87	0.27	0.33	0	0.47
yaho	-0.33	-0.33	-0.47	-0.47	-0.47	0.0	-0.73	-0.60	-0.47	0

Table 7: *Normalised balance values (software)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	-0.61	-0.71	0.18	-0.53	0.30	-0.23	-0.052	0.15	-0.076
bbc	0.61	0	-0.095	0.80	0.082	0.91	0.38	0.56	0.76	0.54
elem	0.71	0.095	0	0.89	0.18	1.0	0.47	0.66	0.85	0.63
empi	-0.18	-0.80	-0.89	0	-0.71	0.12	-0.42	-0.24	-0.036	-0.26
esn	0.53	-0.082	-0.18	0.71	0	0.83	0.3	0.48	0.68	0.45
imdb	-0.30	-0.91	-1.0	-0.12	-0.83	0	-0.53	-0.35	-0.15	-0.38
sncb	0.23	-0.38	-0.47	0.42	-0.3	0.53	0	0.18	0.38	0.16
unde	0.052	-0.56	-0.66	0.24	-0.48	0.35	-0.18	0	0.2	-0.024
virg	-0.15	-0.76	-0.85	0.036	-0.68	0.15	-0.38	-0.2	0	-0.22
yaho	0.076	-0.54	-0.63	0.26	-0.45	0.38	-0.16	0.024	0.22	0

Table 8: *Normalised* horizontal *balance values (software)*

/	airb	bbc	elem	empi	esn	imdb	sncb	unde	virg	yaho
airb	0	-0.58	-0.57	-0.56	-0.46	0.42	-0.45	-0.58	-0.5	0.40
bbc	0.58	0	0.014	0.025	0.12	1.0	0.13	0.0032	0.077	0.99
elem	0.57	-0.014	0	0.011	0.11	0.99	0.11	-0.011	0.063	0.97
empi	0.56	-0.025	-0.011	0	0.095	0.98	0.1	-0.022	0.051	0.96
esn	0.46	-0.12	-0.11	-0.095	0	0.88	0.0059	-0.12	-0.044	0.87
imdb	-0.42	-1.0	-0.99	-0.98	-0.88	0	-0.87	-1.0	-0.92	-0.015
sncb	0.45	-0.13	-0.11	-0.1	-0.0059	0.87	0	-0.12	-0.050	0.86
unde	0.58	-0.0032	0.011	0.022	0.12	1.0	0.12	0	0.073	0.98
virg	0.5	-0.077	-0.063	-0.051	0.044	0.92	0.050	-0.073	0	0.91
yaho	-0.40	-0.99	-0.97	-0.96	-0.87	0.015	-0.86	-0.98	-0.91	0

Table 9: *Aesthetic metrics correlations table*

Aesthetic metric	Pearson	Spearman	Kendall
Balance	0.053	0.040	0.027
Horizontal Balance	0.523	0.527	0.405
Proportion	0.245	0.205	0.140
Simplicity	0.587	0.572	0.424
Symmetry	0.289	0.340	0.226

to judge the UI for its simplicity facets only. This suggests that the algorithm used to compute the simplicity of a UI is quite efficient and aligned with the human judgement.

3.3.4. Balance score and comparison with the software

For the sake of concision, only the results for the technique with the best correlations (simplicity and balance) are reproduced. Full details are publicly available at http://sites.uclouvain.be/questim/bhci2016/. A similar discussion could be provided for the other metrics that are the subject of this paper and the other metrics computed by QUESTIM. Regarding the balance, we proceeded with the same method for comparing scores of balance: obtaining Table 6 for both normalised participants for other preferences and Table 8 representing the software preferences and compare. Colors in Table 8 indicate that the correlation is high between the balance scores rated by participants and those rated by the software. This suggests that the computation metric for balance is reliable enough to be a good predictor of the representation constructed by human beings about the UI balance. Correlation is higher when the measure only focuses on the horizontal components of balance as opposed to balance computed as a whole.

3.3.5. Correlation tables

Table 9 confirms the previous perspective giving three correlation measures for the five considered techniques (including horizontal balance): correlation values between human's balance and software's horizontal balance (in opposition to general balance) and between human's simplicity and software's simplicity are high. Another potentially interesting correlation table is the global Pearson correlations between all data tables in Table 10, that similarly to previous tables, colored the positive (green) and negative (red) values to bring forward relevant elements. The headers in column 1 and row 1 are the considered techniques with (H) if it corresponds to the data gathered among the participants or (Q) if the score is given by the tool. Therefore, we notice that all participants reviews then to be correlated even if they are focused on different visual aspects. The computed metrics of balance and simplicity presents a high level of correlation.

3.4. Discussion

Although the number of participants in the experiment was relatively small, the Randolph's Kappa coefficients indicate a medium degree of inter-judge agreement, which certifies that the participants could somehow reach a consensus. As suggested by the tables examined insofar, the metrics used for computing simplicity and horizontal balance seem to be good predictors while metrics used for symmetry and proportion display a somewhat low reliability, thus suggesting that the computed metrics for these two properties must be improved in order to better represent human perceptions. A possibility to improve the metric for balance could be to add to the algorithm other parts of the dimensions included in the formula proposed by Ngo et al. (2000). Our implementation was limited to objects size and position whereas the formula took also into account objects colors and complexity.

Another possible line of research is to tackle the problem at its roots by questioning the method used by the software to determine UI objects. Region segmentation is based on a semi-automatic process which involves a human action for specifying what and where are the objects in the layout, therefore inducing some degree of probable subjectivity. For instance, a web page menu could be considered as a single region or as a series of headers, labels, and menu items. When these UI elements have their own specific formatting regulated by the CSS, there is a trend to consider them as different UI elements, whereas a same style would tend to consider them as a single piece. This subjectivity could bias the software and prevent it from finding the genuine value

Table 10: *Pearson correlations*

/	AEST(H)	BAL(H)	BAL(Q)	H-BAL(Q)	PROP(H)	PROP(Q)	SIMP(H)	SIMP(Q)	SYMM(H)	SYMM(Q)
AEST(H)	1.0	0.75	-0.01	0.45	0.73	0.14	0.85	0.46	0.74	0.16
BAL(H)	0.75	1.0	0.05	0.52	0.80	-0.041	0.69	0.60	0.84	0.17
BAL(Q)	-0.01	0.05	1.0	0.43	0.05	-0.09	0.03	-0.34	-0.04	-0.08
H-BAL(Q)	0.45	0.52	0.43	1.0	0.48	0.21	0.34	0.14	0.34	0.28
PROP(H)	0.73	0.80	0.05	0.48	1.0	0.25	0.85	0.68	0.75	0.16
PROP(Q)	0.14	-0.041	-0.09	0.21	0.25	1.0	0.22	-0.023	-0.26	-0.37
SIMP(H)	0.85	0.69	0.03	0.34	0.85	0.22	1.0	0.59	0.78	0.084
SIMP(Q)	0.46	0.60	-0.34	0.14	0.68	-0.023	0.59	1.0	0.69	0.48
SYMM(H)	0.74	0.84	-0.04	0.34	0.75	-0.26	0.78	0.69	1.0	0.29
SYMM(Q)	0.16	0.17	-0.08	0.28	0.16	-0.37	0.084	0.48	0.29	1.0

of the balance. It is also likely that participants identified more or less UI elements than the ones spotted by the software. A better strategy would be to analyze interfaces with pixel-based measurement instruments in order to avoid completely to involve a human in the process. Concretely, for the latter approach and as static UIs can be abstracted to simple images, methods for computing balance according to saliency (Shen and Zhao (2014)) should be considered. Furthermore, measuring visual complexity - or simplicity - with the help of the compression level of an image is also meaningful (Tuch et al. (2012)). Another idea could be to use machine learning by confronting a set of UI evaluations to a set of atomic variables characterizing a UI i.e. quantifying objects, colors, dimensions and complexity, in order to find the main characteristics of a well designed UI. But then, we are falling again in the first category of software discussed in 2.3.

4. CONCLUSION

This paper reports on an experiment comparing the results issued from the inter-subjectivity of judgment of fifteen participants evaluating four main aesthetic metrics on a ten graphical user interfaces and the values of these metrics calculated semi-automatically by QUESTIM. QUESTIM is a web-based application that provides Aesthetic Evaluation as-a-Service (AestaaS); which means it is ready to use for any graphical user interface, either directly captured from the web or imported from other formats such as screenshots, wireframes, prototypes, hand-drawn sketches. In this study, ten web pages were captured directly in QUESTIM, the other formats could be considered in a further study to identify whether there is any difference depending on the input format. QUESTIM is a freely accessible service on the web, therefore there is no need for the designer or the developer to install any stand alone application, any plug-in, any add-on or any Java Web Start application. QUESTIM also automatically computes other Ngo and Byrne (2001) metrics, such as equilibrium, cohesion, sequence, unity, regularity, economy, homogeneity, and rhythm. In addition to these metrics, QUESTIM also computes center-alignment, external-alignment, and concentricity, thus making it unique in terms of metric coverage. These metrics could represent a future avenue of this paper.

The experiment suggests that symmetry, proportion, simplicity are positively correlated with human judgment, whereas balance was not. Therefore, a new formula for computing balance was defined that decomposes balance into horizontal and vertical balances, which then revealed some correlation. Another formula for balance, along with its interpretation, is then introduced.

UI aesthetic evaluation is possible based on a limited set of metrics: there exists some positive correlation between the UI simplicity and its overall aesthetics perception, simplicity and balance have high predicting rates when compared to human perceptions of the same metric. Proportion and symmetry only reach a moderate correlation with the participants perceptions, thus calling for other formulas and other interpretations.

In this experiment, only a summative evaluation has been considered: since QUESTIM displays the values of the metrics on top of a colored spectrum, not only the value (summative part), but also its positioning with respect to acceptable range could also be considered (formative part). Any UI element in QUESTIM could be moved, resized, re-aligned, thus enabling the designer to explore alternate designs based on the same elements or not (deleting any UI element is possible, but not yet adding another one). An optimisation problem then arises in order to optimise the global layout under the constraints of metrics being located in acceptable ranges. If solved, QUESTIM would be able to suggest to the designer which operation needs to be performed on which UI element so as to reach an optimum or close by.

5. ACKNOWLEDGMENTS

This work is partially supported by the RW-DGO6 QualIHM project. A special thank goes to students in Human-Computer Interaction course from Université catholique de Louvain (F. Cipollini, M. Bordalo, P. Sanetti and E. La Rocca) for their precious help and support with the experiment.

REFERENCES

Altaboli, A. and Y. Lin (2011, January). Investigating effects of screen layout elements on interface and screen design aesthetics. *Adv. in Hum.-Comp. Int. 2011*, 5:1–5:10.

Ben-Bassat, T., J. Meyer, and N. Tractinsky (2006, June). Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Trans. Comput.-Hum. Interact. 13*(2), 210–234.

Clay, F. (1908). The origin of the aesthetic emotion. *Sammelbande der Int. Musik.*, 282–290.

Comber, T. and J. Maltby (1997). Layout complexity: does it measure usability? In *Human-Computer Interact. INTERACT'97*, pp. 623–626. Springer US.

Courcoux, P. and M. Semenou (1997). Preference data analysis using a paired comparison model. *Food quality and preference 8*(5), 353–358.

Dion, K., E. Berscheid, and E. Walster (1972). What is beautiful is good. *Journal of personality and social psychology 24*(3), 285.

Hamborg, K.-C., J. Hülsmann, and K. Kaspar (2014, January). The interplay between usability and aesthetics: More evidence for the "what is usable is beautiful" notion. *Adv. in Hum.-Comp. Int. 2014*, 15:15–15:15.

Kant, I. (1987). *The Critique of judgment*. Hackett Publishing.

Kurosu, M. and K. Kashimura (1995). Apparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In *Conference companion on Human factors in computing systems*, pp. 292–293. ACM.

Lindgaard, G., G. Fernandes, C. Dudek, and J. Brown (2006, March). Attention web designers: You have 50 milliseconds to make a good first impression! *Behav. Inf. Technol. 25*(2), 115–126.

Mõttus, M., D. Lamas, M. Pajusalu, and R. Torres (2013). The evaluation of interface aesthetics. In *Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation*, MIDI '13, New York, NY, USA, pp. 3:1–3:10. ACM.

Mahajan, R. and B. Shneiderman (1997, Nov). Visual and textual consistency checking tools for graphical user interfaces. *IEEE Transactions on Software Engineering 23*(11), 722–735.

Moshagen, M. and M. T. Thielsch (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies 68*(10), 689 – 709.

Ngo, D. and J. Byrne (2001). Another look at a model for evaluating interface aesthetics. *Int. J. Appl. Math. Comput. Sci. 11*(2), 515–535.

Ngo, D., A. Samsudin, and R. Abdullah (2000). Aesthetic measures for assessing graphic screens. *J. Inf. Sci. Eng 16*, 97–116.

Ngo, D. C. L., L. S. Teo, and J. G. Byrne (2003). Modelling interface aesthetics. *Information Sciences 152*, 25 – 46.

Nielsen, J. and R. Molich (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 249–256. ACM.

Norman, D. A. (2005). *Emotional design: Why we love (or hate) everyday things*. Basic books. Papachristos,

E. and N. Avouris (2011). Are first impressions about websites only related to visual appeal? In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part I*, INTERACT'11, Berlin, Heidelberg, pp. 489–496. Springer-Verlag.

Purchase, H. C., J. Hamer, A. Jamieson, and O. Ryan (2011). Investigating objective measures of web page aesthetics and usability. In *Proceedings of the Twelfth Australasian User Interface Conference - Volume 117*, AUIC '11, Darlinghurst, Australia, Australia, pp. 19–28. Australian Computer Society, Inc.

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Joensuu University Learning and Instruction Symposium*.

Reinecke, K., T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos (2013). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '13*, 2049.

Robins, D. and J. Holmes (2008). Aesthetics and credibility in web site design. *Information Processing & Management 44*(1), 386–399.

Salimun, C., H. C. Purchase, D. R. Simmons, and S. Brewster (2010a). The effect of aesthetically pleasing composition on visual search performance. *Proc. 6th Nord. Conf. Human-Computer Interact. Extending Boundaries - Nord. '10*, 422.

Salimun, C., H. C. Purchase, D. R. Simmons, and S. A. Brewster (2010b). Preference ranking of screen layout principles. In *Proceedings of the 2010 British Computer Society Conference on Human-Computer Interaction, BCS-HCI 2010, Dundee, United Kingdom, 6-10 September 2010*, pp. 81–87.

Schenkman, B. N. and F. U. Jönsson (2000). Aesthetics and preferences of web pages. *Behaviour & Information Technology 19*(5), 367–377.

Sears, a. (1993, July). Layout appropriateness: a metric for evaluating user interface widget layout. *IEEE Trans. Softw. Eng. 19*(7), 707–719.

Shen, C. and Q. Zhao (2014). Webpage saliency. In *Computer Vision–ECCV 2014*, pp. 33–46. Springer.

Sonderegger, A. and J. Sauer (2010, May). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. *Appl. Ergon. 41*(3), 403–10.

Tractinsky, N., a.S Katz, and D. Ikar (2000, December). What is beautiful is usable. *Interact. Comput. 13*(2), 127–145.

Tuch, A. N., E. E. Presslaber, M. Stöcklin, K. Opwis, and J. a. Bargas-Avila (2012, November). The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *Int. J. Hum. Comput. Stud. 70*(11), 794–811.

Vanderdonckt, J. and X. Gillo (1994). Visual techniques for traditional and multimedia layouts. In *Proc. Work. Adv. Vis. interfaces*, pp. 95–104. ACM.

Zain, J. M., M. Tey, and Y. Goh (2011). Probing a Self-Developed Aesthetics Measurement Application (SDA) in Measuring Aesthetics of Mandarin Learning Web Page Interfaces. *8*(1).