Computation of Interface Aesthetics

Aliaksei Miniukovich
University of Trento
Via Sommarive 9, Povo TN, Italy
miniukovich@disi.unitn.it

Antonella De Angeli
University of Trento
Via Sommarive 9, Povo TN, Italy
deangeli@disi.unitn.it

ABSTRACT
People prefer attractive interfaces. Designers strive to outmatch competitors, and create apps and websites that stand out. However, significant expenses on design are unaffordable to small companies; instead, they could adopt automatic tools of interface aesthetics evaluation, a cheaper strategy to good design. This paper describes an important step towards such a tool; it presents eight automatic metrics of graphical user interface (GUI) aesthetics. We tested the metrics in two exploratory studies – on desktop webpages (N = 62) and on iPhone apps (N = 53) – and found them to function on both GUI types and for both immediate (150ms exposure) and deliberate (4s exposure) aesthetics impressions. Our best-fit regression models explained up to 49% of variance in webpage aesthetics and up to 32% (if app genre is considered) of variance in iPhone app aesthetics. These results confirm past results and suggest the metrics are valid and reliable enough to be widely discussed, and possibly, to be embedded in our prospective GUI evaluation tool, tLight.

Author Keywords
Automatic metrics; GUI evaluation; user study; immediate impression; deliberate impression; tLight.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User interfaces – graphical user interfaces (GUI), evaluation/methodology.

INTRODUCTION
There is no doubt visual aesthetics matters in interface design. Surrounded with multiple offers of same-quality services and products, Web and app users have become selective and disregard apps and websites they do not like immediately [12]. A possible way to survive in such an environment includes carefully working out all details of visual design, making the design stand out [33]. However, small companies, start-ups and individual developers often cannot afford hiring a design agency and do their design themselves. In such cases, even well-detailed design guidelines are of limited help, since, to be applied properly, they require extensive training. Concrete GUI evaluation tools could exemplify abstract design guidelines, and drive and substantiate design choices. The tools would be based on specific quality metrics that represent specific GUI design aspects.

In this paper, we extend earlier work [17, 16], and describe and test in two studies eight GUI aesthetics metrics: visual clutter, color range, number of dominant colors, figure-ground contrast, contour congestion, symmetry, and the new metrics of grid quality and white space. We based the metrics on the psychological investigations of what people see as complex and unappealing [23], and HCI investigations of webpage aesthetics [33, 14, 36, 25, 26]. The results of the present two studies replicated the results of past studies [17, 16], which suggests the metrics are solid enough to be presented to the larger CHI audience. In addition to this, we have replicated the phenomenon of consistent and lasting immediate impressions [14, 33] on two types of stimuli (webpages and mobile apps) using a between-subjects experimental design. In the rest of paper, we review related work on aesthetics in HCI and automatic aesthetics measures, and describe the eight GUI aesthetics metrics. We report Study 1, which tested metric performance on webpages, and Study 2, which tested metric performance on iPhone apps. Lastly, we summarize the results and discuss their implications for the automatic evaluation of interface aesthetics.

RELATED WORK
Past attempts [40, 39, 26, 25, 16, 17] to automatically account for visual aesthetics of GUIs consisted of two steps: gathering user scores of GUI aesthetics and matching them against computed scores of a set of automatic metrics. The first step reflects our understanding that beauty lies in the eye of the beholder, and involves conducting either carefully orchestrated in-lab user studies [14, 33, 21, 17] or large-scale crowdsourcing studies with thousands of participants [25, 25]. The second step uses the averaged user scores as the ground truth data, and tests how well various metrics and algorithms predict the scores.

Collecting Aesthetics Scores
The influence of aesthetics on the overall appreciation of GUI changes with time. Sonderegger et al. [29] conducted a longitudinal study and demonstrated the positive effect of aesthetics to almost disappear after the initial use phase. However, it is largely the initial phase that determines if a

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea
 ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702575

one-time visitor converts to a user or goes to competitors [12]. While considering the initial use phase itself, aesthetics impressions could be further subdivided into the *immediate-first* (formed at a glance), *deliberate-first* (long enough for reading titles and processing images) and *overall* (after performing several tasks) impressions [30].

Several HCI studies have explored the first impression of GUI aesthetics and possible ways of collecting user scores of such impressions. In within-subjects experiments, Lindgaard et al. [14] and Tractinsky et al. [33] showed webpage screenshots to participants for half-second intervals, and asked to set a single rating per screenshot. Their experimental design allowed each participant to rate in a relatively short time a large number of stimuli, which, in turn, allowed the authors to make generalizable inferences about stimuli. Tractinsky et al. [33] then compared the half-second, immediate-first ratings with the ten-second, deliberate-first ratings of each participant and found them to strongly correlate.

Studies [13, 30] have compared the deliberate-first and overall aesthetics impressions, and their outcomes. Lee et al. [13] studied the user preferences of four websites before (deliberate-first impression) and after (overall impression) actual use, and found actual use to significantly change aesthetics appreciation. De Angeli et al. [6] explored the interplay of various qualities of two website GUIs differing in their level of aesthetics. They asked participants to perform tasks and carefully documented perceived usability, aesthetics and information quality. The effect of aesthetics in GUI appreciation was significant. Notably, both studies [6] [13] employed multi-item questionnaires for measuring aesthetics, which, one could argue, is more valid than the one-question approach. However, the whole procedure took several hours for each participant and involved studying only two [6] and four [13] GUIs.

Measuring Aesthetics Automatically

Earlier-proposed measures could be generally categorized in *element-based* and *pixel-based*. Element-based measures require knowing the organizational principles (e.g., which GUI elements can contain other elements) and basic elements of GUI (e.g., buttons, links or paragraphs of text). For example, Michailidou et al. [15] counted the number of menus, images, words and links on webpages, and found these numbers to strongly correlate with the user scores of aesthetics and visual complexity of webpages. Harper et al. [10] split each webpage in blocks of 200 × 200 pixels and counted the top-left corners of webpage elements that fell in a block. The distribution of blocks predicted user ranking of 20 webpages with 86% accuracy. Ngo et al. [19] formulated 14 arts-based measures of graphic display aesthetics. They used the positions and sizes of display elements for quantifying balance, equilibrium, symmetry, sequence, cohesion and nine other webpage characteristics. In a later test of these measures, Purchase et al. [22] used the HTML sources of 15 webpages to locate text, image and control (e.g., button) elements, and thus, to compute measure scores. A subset of the measures predicted the user ranks of webpage visual appeal. Lastly, Ivory et al. [11] suggested a number of measures of webpage design quality, which involved counting words, links, images, font types, colors of links and text, text clusters and other webpage aspects. Their metrics predicted if a webpage would be rated very low or very high on visual appeal with 65% accuracy.

Pixel-based measures take GUI screenshots as an input instead of GUI-underlying code. GUI screenshots might represent better what the user sees, which is why pixel-based metrics are considered advantageous to element-based metrics [26]. Zheng et al. [40] segmented webpage screenshots in color-homogeneous blocks to assess webpage symmetry, balance and equilibrium. A part of their measures correlated with user scores of several aesthetics sub-dimensions. Purchase et al. [21] explained nearly 25% of variance in user scores of visual complexity of various images. Their measures included computing the variance in image pixel luminance, the ratio of contour pixels to all pixels, file sizes of images in various file formats, and the number of image colors. In two large-scale online studies, Reinecke et al. [25, 26] explained up to 49% of variance in webpage aesthetics using the number of text and images blocks, mean values of hue, saturation and value of screenshot pixels, proportion of screenshot pixels of a particular color and several other measures. Finally, Wu et al. [39] listed a number of measures, both pixel-based and element-based. The pixel-based measures included decomposing webpages in visual blocks as the first step, and combining the blocks in hierarchical tree-like structures as the second step. Then, the block-average brightness, hue, saturation, texture and colorfulness were computed. These measures, combined with other element-based measures, allowed Wu et al. [39] to classify webpages with a 77% accuracy, relative to users' classifications. Notably, the webpage decomposition algorithm used in [39] was not pixel-based and still required HTML sources of webpages as an input.

Complexity Roots of Aesthetics

Psychology [23] suggested liking of everyday[1] things might arise from the ease of understanding, i.e., from the simplicity of things. Our brain has evolved to like spending less energy processing things; we see simpler objects as more familiar, and thus, safer. One might argue that complexity is only a single component of liking and conscious considerations might matter even more than complexity [1]. However, the pre-conscious perception of complexity is deemed to be universal [23] and is easier to account for than conscious considerations.

Several studies [5, 18] have explored what complexity means in HCI and listed three types of complexity: *visual complexity*, *information design complexity* and *task complexity*. Visual complexity strongly relates to immediate aesthetics perception, and many HCI studies

[1] We would like to stress that we discuss the aesthetics of everyday things, not pieces of art (cf. [32]).

focused on exploring and leveraging this type of complexity. Tuch et al. [35, 36] have asserted a strong link between visual complexity and immediate aesthetics of webpages. Several other studies [39, 25, 26] tried to measure webpage aesthetics automatically, using quantifiable aspects of HTML sources or screenshots of webpages. Despite these studies started from an assumption of a strong complexity-aesthetics link, none of them looked systematically into the underlying determinants of visual complexity, and thus, missed out several potential determinants from their analyses. Earlier work [16, 17] gave a more detailed overview of visual complexity determinants and listed eight of them, relevant to HCI (Table 1).

Visual complexity determinants	**Status**
Visual clutter	A multi-item metric[a]
Color variability	Two multi-item metrics[a] (dominant colors and color range).
Contour congestion	A single-item metric[a]
Figure-ground contrast	A single-item metric[a]
Layout quality	A multi-item metric[b] (grid quality) and single-item metric[b] (white space)
Symmetry	A single-item metric[c]
Prototypicality	*Not implemented*
Ease of grouping	*Not implemented*

[a] described in [16, 17]; [b] introduced in this paper; [c] described in [16, 17], but fully reworked here.

Table 1. The determinants of visual complexity and associated metrics.

METRICS

The metrics of clutter, color range and dominant colors have evolved from the constructs of visual clutter and color variability. Both constructs describe the amount of information on a screen; however, color variability is often measured and considered separately from clutter due to its salience: study participants always mention colors as a component of complexity [20]. Visual clutter describes the effort to introduce a new, visually prominent object to a scene [26] and is quantified with several measures (CL1-CL4, Table 2). Color variability (measures CV1-CV5, Table 2) consists of two aspects that humans perceive separately [16]: number of dominant colors (the colors a human can easily differentiate and name) and color range (the colors a human cannot differentiate without zooming in, and which are often used for smoothing edges and color gradients).

The metrics of figure-ground contrast and contour congestion reflect the constraints of the human visual system. Figure-ground contrast describes differences in luminance or color of adjacent lines. Smaller differences lead to higher mental effort needed to recognize objects or to read text [9]. As in [16, 17], we operationalized contrast as a weighted sum of number of contour pixels detected with the Canny algorithm at several consecutive levels. The Canny edge detection algorithm takes two input thresholds: high and low. In the present metric implementation, the low threshold is set to 40% of the high threshold; the high threshold varies from 10% to 70% of maximal pixel luminance (5% to 65% for the mobile app screenshots) with a step of 10%. The weights decrease from 1 to 0 with a step of .2; higher weights are assigned to edge pixels detected with lower thresholds. Finally, the weighted sum is normalized by the number of all detected edge pixels and taken as the metric of contrast.

#	**Measure description**
CL1	Contour density, the ratio of contour pixels to all pixels, cf. [27]
CL2	Subband Entropy [27], the amount of redundancy introduced to a scene.
CL3	Feature Congestion [27], the proportion of unused feature (e.g., color or luminance) variance.
CL4	The file size of images in the JPEG format (cf. [35, 27])
CV1	The file size of images in the PNG format.
CV2	The number of colors after color reduction: only values that occurred more than 5 (for web) or 2 (for mobile) times per image were counted.
CV3	The number of colors per dynamic cluster (see CV5).
CV4	The number of static 32-sized color clusters (the sub-cube edge size of clusters is 32 values out of possible 256, per each RGB channel). Only clusters containing more than 5 values are counted.
CV5	The number of dynamic clusters of colors after color reduction (more than 5 pixels). If a difference between two colors in a color cube is less than or equal to 3, two colors are united in the same cluster, which continues recursively for all colors. Only clusters containing more than 5 values are counted.

Table 2. The measures of visual clutter (CL1-4) and color variability (CV1-5).

Contour congestion describes the mental effort needed to differentiate spatially proximal lines. If two objects are too close to each other, a human cannot differentiate them with the peripheral vision and needs to focus on the objects [37]. As in [16, 17], we operationalized contour congestion as the proportion of congested contours to all contours. First, contour pixels are detected. Then, all contour pixels that have neighbors in a 20-pixel vicinity are marked as congested. Finally, the congested pixels are counted and normalized by the number of all edge pixels.

Well-organized information requires less cognitive effort to process. Symmetry [34, 3] and regular visual layout [2] might serve for such a purpose in HCI. In this paper, we reconsidered the past algorithm of GUI symmetry [17, 16] (in which contour symmetry was measured by looking for a match for each contour pixel across the central vertical axis). The past measure was too noisy and favored GUIs with fewer objects. Here, we measured block symmetry, which considers the position of GUI visual blocks, relative to the central vertical axis. First, we partitioned a GUI screenshot into visual blocks (the algorithm was inspired by [4]). Second, we considered separately the blocks that contained the central vertical axis and blocks that did not.

The shift of the former relative to the axis was considered as asymmetry. The shift of the latter relative to the axis and matching block (if there was a matching block) was also considered as asymmetry.

The grid quality and white space metrics describe the quality of GUI layout. Higher quality helps the user to quickly navigate within the GUI and is seen as an important aesthetic aspect of GUI [2]. We considered several existing measures of alignment and regularity of document layouts [2], and implemented those that did not require a high precision detector of GUI block positions. We first sliced a GUI screenshots in visual blocks. We then assumed the non-covered proportion of screenshot to reflect badly distributed content and took it as the white space metric. The other five block-based measures (Table 3) describe grid quality and can be combined using a factor analysis. The last measure (Table 3, G5) was added specifically to reflect the specificity of mobile GUIs: they are often organized in a single aligned column of elements, which users are tolerant of scrolling down through. Thus, the measures G1 and G2 should not apply to mobile GUIs.

#	Measure description
G1	The number of visual blocks of GUI (cf. [2]).
G2	The number of alignment points of blocks (cf. [2]).
G3	The number of block sizes, grid proportionality [2].
G4	The proportion of GUI covered by same-size blocks (cf. the cell coverage computation from [2]).
G5	The number of vertical block sizes, i.e., vertical grid proportionality.

Table 3. The measures of GUI grid quality.

STUDY 1

The first study sought to replicate the past results [16] on a bigger and more diverse sample of websites and pool of participants. In addition, we strived to extend the results from *immediate-first* to *deliberate-first* aesthetics impressions [30]. Past studies of immediate-deliberate differences [38, 33] used a within-subjects experimental design, whereas we used a between-subjects design: half of participants rated screenshots after a 150ms exposure and the other half after a 4s exposure. After we collected user ratings of aesthetics of 300 webpages, we matched the mean user scores against the scores of the automatic metrics. Additionally, we applied the resulting regression model of aesthetics to stimuli from [16].

Stimuli

A representative sample of stimuli largely determines the validity of findings. Researchers either selected websites themselves [35], relied on online collections of websites preselected for their high quality [40, 16], or asked design professionals to submit website links [14]. All these approaches implied a sample bias, either because of idiosyncratic preferences of the few people involved in the selection or because of a possible experimenter effect whereby researchers could unconsciously select the "right" stimuli. We, instead, outsourced website search and selection to a larger number of people via crowdsourcing. Over 40 workers of a crowdsourcing platform[2] looked for and reported websites in English of three website genres: corporate, eCommerce and news. We then reviewed over 300 submitted websites and filtered out those that did not match the genres we required, did not have all the page types we required (Table 4), had technical issues or required to login before displaying main content. Keeping in mind that familiarity could seriously complicate data analyses [31], we also avoided the top 500 popular websites[3] and their localized versions (e.g., amazon.it instead of amazon.com). The filtered set contained 235 websites from which we randomly selected 75 websites, 25 in each genre. Finally, we automatically took 300 screenshots of website pages (1280×800 pixels, webpage top part only, PNG format, 24-bit per pixel).

Genre	Corporate	Ecommerce	News
Page types	home	home	home
	about us	about us	about us
	contact us	details of item	piece of news
	products & services	list of items	list of news

Table 4. We took four genre-specific pages of each website.

Participants

We recruited 62 participants (mean age = 31.4 years, SD = 6.3; 22 female; 30 non Italians from all over the globe) including 7 students, 27 doctoral students, 11 postdocs and 17 full-time university employees; all were proficient in English and had normal or corrected to normal vision. Participants reported spending 6.3h a day on the Internet (SD = 3.4h); 40 participants had technical background; 21 participants indicated having significant experience in visual or GUI design. One person did not finish the test.

Design

We adopted a one-way between-subjects experimental design with exposure duration (150ms vs. 4s) as an independent factor and visual aesthetics as a continuous dependent variable. Participants were randomly assigned to either the 150ms or 4s condition. This manipulation should discern *immediate-first* from *deliberate-first* impressions: 150ms is only long enough to grasp the gist of scene [7] but not enough for reading [28], whereas 4s is long enough for up to 10 eye fixations and reading headlines. The standard duration of 500ms [33, 14] was discarded as it would allow for 1-2 eye fixations and reading 1-2 words, and would mix immediate with deliberate impressions.

Procedure

All test sessions were conducted individually, in an isolated room with a laptop and experimenter, and started with a briefing form and consent form. Participants then filled out demographics questionnaires. The study consisted of viewing (1280×800, 13-inch display) and rating 100 webpage screenshots, randomly selected from

[2] https://microworkers.com

[3] http://www.alexa.com/topsites

the pool of 300 screenshots. In each trial (cf. [16]), participants saw a fixation cross (1-1.5sec), a webpage screenshot (150ms or 4s, depending on a participant condition), and black-white noise screen (50ms) and were prompted to rate the screenshot with the key buttons from one (ugly) to seven (beautiful). We did not limit the time for rating but explicitly asked participants to do it as quickly as possible. The average completion time was below 15 minutes. At the end of test, participants were debriefed and given a small reward.

Results

In the debriefing phase, participants often noted they disliked the "broken" shopping cart pages of eCommerce websites. (Almost all shopping cart pages featured a message that the shopping cart was empty, often with the inclusion of such words as "sorry" or "unfortunately".) We did not aim at accounting for this emotional bias and decided to exclude shopping cart pages from the analysis. The ratings of one participant assigned to the 150ms condition (mean = 6.02, on a 1 to 7 scale) deviated from the mean of others by nearly three standard deviations and were excluded from further analyses. Thus, we obtained 9 to 11 ratings per screenshot.

The average score interclass correlation coefficients suggested a high consistency in user scores in both 150ms (ICC2k = .77; 95% conf. interval is .74 to 0.81; $F(274, 8970) = 4.95$, $p < .001$) and 4s (ICC2k = .85; 95% conf. interval is 0.83 to 0.88; $F(274, 8671) = 7.08$, $p < .001$) conditions. The analysis of mean scores (from now on we refer to means per screenshot) indicated our sample included both appealing and non-appealing webpages (ranging from 2.1 to 6.1 on a 1 to 7 scale, Figure 1) and was only slightly skewed in both 150ms (n = 275, mean = 3.79, SD = .8, min = 2.1, max = 6.1, skew = .48) and 4s (n = 275, mean = 3.9, SD = .89, min = 1.9, max = 6.3, skew = .24) conditions. A paired t-test revealed a small, but significant difference between the mean scores of 150ms and 4s conditions, diff. = .11, $t(274) = 2.82$, $p < .01$. The correlation between the mean scores of 150ms and 4s conditions was strong, $r(273) = .70$, $p < .001$.

We computed scores of the eight metrics for each screenshot. Four metrics (visual clutter, number of dominant colors, color range and grid quality) required combining multiple measures, for which we used maximum-likelihood factorial analyses. As in [17, 16], we expected the variance of visual clutter and color variability measures to partially overlap and wanted to cancel out the overlapping – we analyzed all the measures in a factor analysis with Varimax rotation (Table 5a). Three factors emerged. Similar to [17, 16], no measures had high cross loadings, with the exception of number of static color clusters (CV4, Table 2), which loaded on both clutter and dominant color factors. We computed Thompson's scores for the three factors. The measures of grid quality (G1-G4, Table 3) were combined in a separate factor analysis (Table 5b), in which a single factor emerged. We again estimated grid quality scores using Thompson's method.

Figure 1. The least appealing (top, 2.1 out of 7) and the most appealing (bottom, 6.1 out of 7) webpages.

	measures / factors	CL1	CL2	CL3	CL4	CV1	CV2	CV3	CV4	CV5
A	Clutter	.92	.85	.95	.92	.30		-.14	.51	.30
	Color range		.20		.18	.82	.89	.72	.13	
	Dominant colors	.26	.29	.24	.22	.15	.21	-.36	.65	.85

	measures / factors	G1	G2	G3	G4
B	Grid quality	.77	.67	.90	-.87

Table 5. Factor loadings of clutter and color measures (A, cumulative var. g = .83) and grid quality measures (B, cumulative var. g = .65) for webpages.

The majority of the automatic metrics generated scores that correlated with the user scores (Table 6). Please note that higher symmetry and grid quality scores correspond to less symmetric and less organized layouts, and therefore, correlate negatively with aesthetics. We then used the Akaike's Information Criterion (AIC) to select the best-fit linear models, which explained 49% (150ms condition) and 43% (4s condition) of user score variance. AIC, in addition to looking at R^2, penalizes models for each additional predictor.

Finally, we applied the factor loadings (Table 5) and linear model coefficients (Table 7, 150ms condition) from this study to the stimuli from [16], i.e., to a different data set. The amount of explained variance in the aesthetics ratings only dropped down to 42% (from 51% in [16]).

Discussion

Consistent with the past results [16], the automatic aesthetics metrics indeed captured certain aspects of webpage visual aesthetics, with clutter and color range

being the strongest predictors. Two new metrics (grid quality and white space) also performed fairly well and generated scores that correlated with aesthetics ratings. Our best-fit linear regression model accounted for 49% of variance in the ratings of immediate-first aesthetics, which is comparable to the past results for similar stimuli [16]. Moreover, when applied to a different data set (from [16]), the model performed well and explained 42% of rating variance, despite significant differences in data collection of the present and former studies. Notably, we consider the factor loadings (Table 5a) as more valid, compared with the past efforts [16, 17], as they are based on a larger and more diverse sample of screenshots.

Metric	150ms exposure		4s exposure	
	r(273)	*p*	*r(273)*	*p*
Clutter	-.30	< .001	-.48	< .001
Color range	.56	< .001	.33	< .001
Dominant colors	-.08	.14	-.20	< .01
Contrast	.51	< .001	.37	< .001
Congestion	-.46	< .001	-.37	< .001
Symmetry	-.16	< .01	-.11	.07
Grid quality	-.22	< .001	-.22	< .001
White space	-.15	< .05	-.01	.90

Table 6. Pearson's correlation coefficients between metric-produced scores and user scores.

	150ms exposure		4s exposure	
Predictor	*Estimate*	*p*	*Estimate*	*p*
(Intercept)	3.79	< .001	3.90	< .001
Color range	.34	< .001	.23	< .001
Clutter	-.21	< .001	-.51	< .001
Dominant colors	-.11	< .01	-.24	< .001
Grid quality	.15	< .01	.17	< .01
Contrast	.12	< .05	--	--
Symmetry	-.14	< .01	-.08	.12
Congestion	-.12	< .01	-.07	.14
White space	-.08	.12	-.20	< .01
R^2 (adj. R^2)	.49* (.48)		.43** (.42)	

* $F(8,266) = 32.52$; ** $F(7,267) = 29.17$; both $p < .001$.

Table 7. Regression models of webpage visual aesthetics. (Outcome variables are aesthetics scores after 150ms and 4s exposure).

Although we cannot directly compare the results (Table 7) with the results of other similar studies [25, 39], we would like to point out that the metrics of [39], despite an impressive 77% accuracy in predicting user ranking of webpage complexity, performed only 7% better than the Feature Congestion complexity measure [27]. Feature Congestion was only one of many measures we integrated in the metrics and when we tried to use it alone to explain aesthetics ratings, R2 dropped to .11 (from .49, Table 7). The webpage aesthetics models of [25] performed fairly well (R^2 = .47), but they included several culture-dependent predictors (e.g., distaste for a particular color) and demographic predictors (age, gender, geographic location and education level), whereas we used none of them. Collecting demographics and preference data increase GUI evaluation time and cost, and makes an evaluation less affordable to small companies.

Our experimental procedure used a between-subjects experimental design, in which we collected ratings of both *immediate-first* (150ms) and *elaborate-first* (4s) impression of aesthetics. A paired t-test showed elaborate impressions to be slightly more positive than immediate impression (diff. = .11, $p < .01$), which was expected (cf. [24]) and could reflect a lower cognitive strain after watching screenshots for longer time [33]. The correlation between *immediate-first* and *deliberate-first* scores was strong, $r(298) = .70$, $p < .001$, which supported earlier inferences of within-subjects-design studies [33, 14]: participants do form an impression almost instantly and this impression lasts. In fact, several participants of the study explicitly mentioned that the presentation time (4s) could have been shorter and they would still rate screenshots the same. When we used the scores of immediate and elaborate impressions in regression analyses, the R^2 of the resulting models (Table 7) only differed by 6%. This seems to suggest that people rely on similar GUI aspects in judging beauty, regardless of exposure duration.

STUDY 2

A previous study [17] argued that pixel-based complexity metrics could successfully evaluate any type of GUIs, not only webpages. Study 2 sought to support the argument and replicate past studies ([17], study 1 and 2) on a bigger and more diverse sample of stimuli (iPhone apps) and pool of participants. The design of Study 2 resembled the design of Study 1; hence we only highlight the differences between the two.

Stimuli

We did not outsource stimuli selection to crowd workers due to several practical considerations. First, we could not find a service for automatic iPhone app GUI rendering and screenshot capturing, and therefore, this had to be done manually. Second, if we asked crowd workers to send us screenshots, we would receive screenshots of different sizes due to differences between various iPhone devices. In addition, we could run into privacy issues due to the possibility of screenshots containing sensitive user data. Finally, crowd tasks are designed to be short and concise, whereas we would set multiple requirements and conditions. This would make task descriptions impractically lengthy.

Instead, we crawled the Apple's app store website and selected free apps in English from three genres: business, travel and entertainment. These genres should represent well the whole of the task-fun continuum, with business apps being mainly task-oriented, entertainment apps being mainly fun-oriented, and travel apps occupying the middle. After we selected three lists of apps, we randomized the app order and considered the apps one by one from the top according to several criteria. First, we only considered apps designed for the 4-inch displays (the apps for smaller

displays would render with two black areas at the top and bottom of display). Second, we avoided overly simplistic apps that had less than four visually diverse GUI layouts. Third, we avoided apps with the heavy use of cartoon graphics. Forth, we avoided apps that required a paid premium account to access their full functionality. Last, to reduce possible familiarity effects, we avoided the apps from the top 100 most popular apps list. After randomly selecting 25 apps per genre, we manually took four or more screenshots (640 × 1136 pixels; PNG format, 24-bit per pixel) per app and then randomly reduced to only four screenshots per app. Thus, we collected 300 screenshots of 75 iPhone apps.

Participants

We recruited 53 participants (mean age = 29.6 years, SD = 7.1 years; 18 female; 16 non-Italians), including 18 students, 17 doctoral students, 6 postdocs and 12 full-time university employees. All participants were proficient in English, had normal or corrected to normal vision (except one, one-eye-blind participant, whose data were later discarded) and no color blindness. All but 10 participants were very familiar with smartphones and 16 were iPhone users. In total, 34 participants had a technical background; 42 participants indicated having no significant experience in visual or GUI design.

Design & Procedure

The experimental design and procedure mirrored Study 1, with the exception of the stimuli and test device (iPhone 5C); 27 participants were assigned to the 150ms condition and 27 to the 4s condition.

Results

We excluded the data from the participant, who did not have normal or corrected to normal vision, and another participant who visibly paid no attention to the task. Each participant rated 100 randomly selected screenshots out of the pool of 300 screenshots, which resulted in seven to nine ratings per screenshot.

The average score interclass correlation coefficients suggested an acceptable consistency in user scores in the 150ms condition (ICC2k = .64; 95% conf. interval is .58 to 0.69; $F(299, 7176) = 2.91$, $p < .001$) and high consistency in the 4s condition (ICC2k = .76; 95% conf. interval is 0.72 to 0.80; $F(299, 7475) = 4.38$, $p < .001$). Mean scores (from now on we describe per screenshot means) indicated the sample of apps was not skewed and included both appealing and non-appealing apps in both 150ms (n = 300, mean = 3.71, SD = .71, min = 1.76, max = 5.81, skew = -.07) and 4s (n = 300, mean = 3.88, SD = .83, min = 1.67, max = 6.33, skew = -.05) conditions. A paired t-test suggested a small but significant difference between the mean scores of 4s and 150ms conditions, diff. = .17, $t(299) = 4.23$, $p < .001$. The mean user scores of 150ms and 4s conditions correlated strongly, $r(298) = .59$, $p < .001$.

As in Study 1, we computed scores for the eight metrics, four of which (visual clutter, number of dominant colors, color range and grid quality) needed additional maximum-likelihood factorial analyses to combine multiple measures. Three factors emerged in the factor analysis with Varimax rotation of clutter and color measures (Table 8a). One factor emerged (Table 8b) in the factor analysis of the grid quality measures relevant to mobile GUIs (G3-G5, Table 3). We then estimated the scores of corresponding metrics using Thompson's method.

	factors \ measures	CL1	CL2	CL3	CL4	CV1	CV2	CV3	CV4	CV5
A	Clutter	.94	.86	.94	.89	.19	.14		.22	
	Color range	.10	.17	-.11	.31	.91	.76	.70	.22	
	Dominant colors		.23		.15	.35	.43		.69	.99

	factors \ measures	G3	G4	G5
B	Grid quality	.99	-.64	.86

Table 8. Factor loadings of clutter and color measures (A, cumulative var. g = .82) and grid quality measures (B, cumulative var. g = .72) for iPhone apps.

The output of the symmetry metric was not normally distributed, since a large part of app layouts were almost perfectly symmetrical. We converted the output of the symmetry metric in a categorical variable: screenshots with a score < 10 were considered symmetrical (104 app layouts); screenshots with a score > 200 were considered asymmetrical (90 app layouts); the rest was considered as partially symmetrical (106 app layouts). Screenshots of the same app often varied widely on symmetry (Figure 2). A one-way ANOVA test revealed a significant effect of symmetry on user score in 4s condition ($F(2,297) = 4.62$, $p < .05$), but not in the 150ms condition ($F(2,297) = 1.01$, $p = .37$). Post-hoc comparisons using the Tukey HSD test showed participants disliked partially symmetrical layouts relative to fully symmetrical and asymmetrical layouts (Table 9b). For the rest of metrics, we estimated the correlations between their output and user scores (Table 9a). The best-fit linear models (based on the Akaike's Information Criterion) explained 13% (150ms condition) and 18% (4s condition) of user score variance (Table 10). When the effect of app genre was considered, the fit of the models went up (17% in the 150ms condition and 32% in the 4s condition).

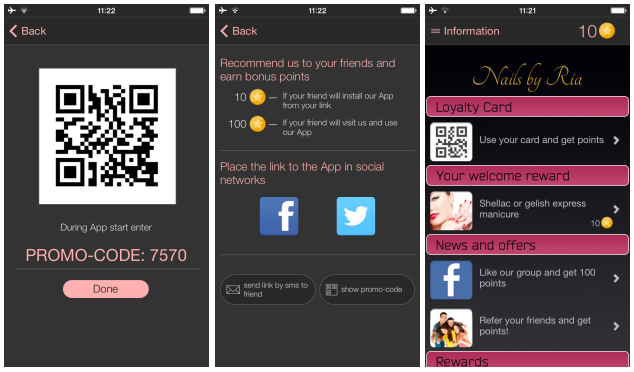

Figure 2. Symmetrical (left), partially symmetrical (center) and asymmetrical (right) screen layouts of an app.

	Metrics	150ms exposure		4s exposure	
		r(298)	*p*	*r(298)*	*p*
A	Clutter	-.22	< .001	-.30	< .001
	Color range	.26	< .001	.28	< .001
	Dominant colors	.09	.13	-.08	.18
	Contrast	.24	< .001	.26	< .001
	Congestion	-.18	< .01	-.22	< .001
	Grid quality	-.13	< .05	-.10	.07
	White space	-.01	.90	.01	.90
B	Symmetry levels	*diff.*	*p adj.*	*diff.*	*p adj.*
	Symm. – Part.Symm.	.18	.16	.32	< .05
	Symm. – Asymm.	.12	.47	.05	.92
	Part.Symm – Asymm.	- .06	.83	- .27	.05

Table 9. Pearson's correlation coefficients of user scores and continuous metrics, and between-level differences for the categorical metric of symmetry.

	150ms exposure		4s exposure	
Predictor	*Estimate*	*p*	*Estimate*	*p*
(Intercept)	3.70	< .001	3.96	< .001
Factor: part. symmetry[#]	--	--	- .20	.07
Factor: full symmetry[#]	--	--	- .02	.84
Clutter	-.13	< .01	-.20	< .001
Color range	.16	< .001	.19	< .001
Contrast	.08	.06	.05	.09
R^2 (adj. R^2)	.13[*] (.12)		.18 (.17)	
R^2 (app genre is included)	.17		.32	

[#] The reference value is complete asymmetry;[*] $F(3,296) = 14.32$,[**] $F(5, 294) = 13.2$, both $p < .001$.

Table 10. Linear regression models of app visual aesthetics. (Outcomes are aesthetics scores after 150ms or 4s exposure).

We also applied the factor loadings (Table 8) and linear model coefficients (Table 10, 150ms condition) from this study to the Android app screenshots from [17]. The amount of explained variance in the Android app aesthetics only dropped down to 30% relative to 36% in [17].

Discussion

Overall, the results of study 2 confirmed our expectations: the majority of the automatic metrics generated scores correlating with the user ratings of iPhone app aesthetics (Table 9). Compared with a similar study on Android apps [17], the best-fit models (Table 10) were less effective at explaining iPhone app aesthetics. However, when applied to the past dataset [17], the present study model (Table 10, the 150ms condition) explained 30% of Android app aesthetics, which demonstrated the reliability of the metrics and suggested they can be applied to both web and mobile GUIs. We consider the factor loadings (Table 8) as more valid compared with our past efforts [17], as they are based on a larger and more diverse sample of screenshots. The inclusion of app genre in the models increases R^2 (up to .32 in the 4s condition), but significantly complicates the analysis. Further studies are needed.

Study 2 used between-subjects experimental design, comparing the ratings of both *immediate-first* (150ms) and *deliberate-first* (4s) impressions of app beauty. As in Study 1, a paired t-test showed elaborate impressions to be slightly more positive than immediate impressions (diff. = .17, $p < .001$), which was also consistent with the past research on affective judgments [24, study 3]. However, the correlation between immediate and deliberate impression ratings remained strong ($r = .58$, $p < .001$). This further supports the claims of aesthetics impressions forming immediately and lasting [33, 14] and extends them from website-only to mobile GUIs as well.

GENERAL DISCUSSION

Study 1 and Study 2 largely converge. Using between-subjects experimental design, both studies observed a strong link between *immediate-first* and *deliberate-first* aesthetics impressions (Table 11a). This reinforces past findings [38, 33] and extends them from the web to the mobile domain (see also [17], study 1).

Passing from immediate-first to deliberate-first impression brings us closer to the real-world usage situations. (One might even argue immediate impressions are only possible in a lab.) The automatic metrics were initially designed to account only for low-level, perceptual GUI qualities, and not for high-level conscious elaborations. However, reading titles and considering images do not drastically change user scores (Table 11a) and performance of the metrics (Tables 6 & 9). We assume that the influence of GUI visual aspects carries over from the very initial, 150ms-long phase to a more elaborate, 4s-long phase.

			Web	Mobile
A	**4s VS 150ms**	Diff. in means	.11[**]	.17[***]
		Correlation	.70[***]	.58[***]
B	**Corr. with 150ms user scores**	Clutter	-.30[***]	-.22[***]
		Color range	.56[***]	.26[***]
		Dominant colors	---	---
		Contrast	.51[***]	.24[***]
		Congestion	-.46[***]	-.18[**]
		Symmetry	-.16[**]	---
		Grid quality	-.22[***]	-.13[*]
		White space	-.15[*]	---
C	R^2 of best-fit 150ms(4s) model		.49 (.43)	.13 (.18)

[***] $p < .001$; [**] $p < .01$; [*] $p < .05$.

Table 11. Comparison of results of studies on web and mobile GUIs.

Predicting GUI aesthetics has proven to be harder for mobile apps than for webpages, which the comparison of R^2 of regression models demonstrates clearly (Table 11c). We could attribute this to two reasons. First, visual complexity might matter much less to mobile app than to webpage evaluators, either because all apps are already designed with complexity concerns in mind (e.g., for the on-the-go use or difficult lighting conditions) or because smaller screens imply fewer details to perceive. Thus, the complexity-rooted metrics explain less of mobile aesthetics. Second, Apple reviews all iPhone apps published on their app store, which means the apps are already preselected based, in part, on their design. The iOS

design guidelines[4] advices developers to "*use a family of pure, clean system colors that look good at every tint …*", "*align text, images, and buttons to show users how information is related"* or to "*create a layout that fits the screen of an iOS device*." Following the guidelines restricts the variance in the scores of several metrics, which also reduces the chances for covariance, i.e. for stronger correlations. Indeed, the metric performance triples on Android apps (R^2 = .30), which Google Play Store publishes without review. We call for more studies on mobile apps, as they are becoming increasingly more important, and yet, rarely mentioned in literature.

The web-mobile comparison of metric performance (Table 11b) also revealed significant resemblance. Except white space, symmetry and dominant color, all metrics functioned similarly (in both, correlation direction and magnitude) for both web and mobile GUIs. The effect of dominant colors did not reach a significant level, which could reflect the overall tendency to use fewer dominant colors in mobile apps: restricted variance in dominant colors would lead to limited chances for covariance with aesthetics. The white space metric did not apply to mobile GUIs, which could follow from participants tolerating well incomplete list- or menu-like GUIs of many mobile apps. Finally, the output of symmetry metric for iPhone apps was not normally distributed (one-column layouts of apps often imply full symmetry) and was converted to a categorical variable with three levels (full symmetry, partial symmetry, and complete asymmetry). Instead of a linear drop in liking from full symmetry to complete asymmetry, we observed partial symmetry to be disliked the most (Table 9b), which was in part consistent with the recent research on slight asymmetries [8]. This might also mean complete asymmetry was perceived as an intended feature and tolerated. Overall, we conclude that the same factors matter for both web and mobile GUI aesthetics, and accordingly, the metric can be similarly applied to both web and mobile GUIs.

CONCLUSION

This paper presented two validation studies of eight automatic metrics of GUI aesthetics. The metrics performed fairly well for websites (Study 1), but were more problematic for mobile apps (Study 2). Each metric accounted for a unique GUI design aspect and could be translated in a design guideline. This work has advanced us towards the final goal of implementing the metrics in tLight, a software tool for helping non-professional designers in creating more appealing and competitive GUIs, and speeding up GUI development cycles. However, reaching the final goal requires several more steps: considering the effect of website or app genre on aesthetics; testing the metrics on aesthetics ratings gathered in more realistic usage contexts, and possibly, with users less technically literate than in the current studies; and establishing the link between approach-avoidance tendencies (e.g., buying or recommending) and the predictions of the metrics. Lastly, tLight (and similar systems) may not completely replace user studies, but can effectively complement them. Studying how user free-form feedback on designs combines with the tLight output would finalize the cycle of tLight-related research.

[4] https://developer.apple.com/design/tips

REFERENCES

1. Armstrong, T. and Detweiler-Bedel, B. Beauty as an emotion: the exhilarating prospect of mastering a challenging world. *Review of general psychology*, 12, 4 (2008), 305-329.
2. Balinsky, H. Evaluating interface aesthetics: measure of symmetry. In *Digital Publishing Conference* (San Jose 2006), International Society for Optics and Photonics.
3. Balinsky, H. Y., Wiley, A. J., & Roberts, M. C. Aesthetic measure of alignment and regularity. In *the 9th ACM symposium on Document Engineering* (Munich 2009), ACM, 56-65.
4. Cao, J., Mao, B., and Luo, J. A segmentation method for web page analysis using shrinking and dividing. *International Journal of Parallel, Emergent and Distributed Systems*, 25, 2 (2010), 93-104.
5. Choi, J. H. and Lee, H. J. Facets of simplicity for the smartphone interface: A structural model. *International Journal of Human-Computer Studies*, 70, 2 (2012), 129-142.
6. De Angeli, A., Sutcliffe, A., and Hartmann, J. Interaction, usability and aesthetics: what influences users' preferences? In *the 6th conference on Designing Interactive systems* (2006), ACM, 271-280.
7. Fei-Fei, L., Iyer, A., Koch, C., and Perona, P. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7, 1 (2007), 1-29.
8. Gartus, A. and Leder, H. The small step toward asymmetry: Aesthetic judgment of broken symmetries. *i-Perception*, 4, 5 (2013), 352-355.
9. Hall, R. H. and Hanna, P. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23, 3 (2004), 183-195.
10. Harper, S., Michailidou, E., and Stevens, R. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception*, 6, 2 (2009).
11. Ivory, M. Y., Sinha, R. R., and Hearst, M. A. Empirically validated web page design metrics. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2001), ACM, 53-60.
12. Kim, H. and Fesenmaier, D. R. Persuasive design of destination web sites: An analysis of first impression. *Journal of Travel Research.*, 47, 1 (2008), 3-13.
13. Lee, S. and Koubek, R. J. Understanding user preferences based on usability and aesthetics before

and after actual use. *Interacting with Computers*, 22, 6 (2010), 530-543.

14. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention web designers: You have 50 milliseconds to make a good first impressoin! *Behavior & information technology*, 25, 2 (2006), 115-126.
15. Michailidou, E., Harper, S., and Bechhofer, S. Visual complexity and aesthetic perception of web pages. In *the 26th annual ACM international conference on Design of communication* (2008), ACM, 215-224.
16. Miniukovich, A., De Angeli A.. Quantification of Interface Visual Complexity. In *the 2014 International Working Conference on Advanced Visual Interfaces* (2014), ACM, 153-160.
17. Miniukovich, A., De Angeli A.. Visual Impression of Mobile App Interfaces. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (2014), ACM, 31-40.
18. Nadkarni, S. and Gupta, R. A Task-Based Model of Perceived Website Complexity. *Mis Quarterly*, 31, 3 (2007).
19. Ngo, D. C. L., Teo, L. S., and Byrne, J. G. Modelling interface aesthetics. *Information Sciences*, 152 (2003), 25-46.
20. Oliva, A., Mack, M. L., Shrestha, M., and Peeper, A. Identifying the perceptual dimensions of visual complexity of scenes. In *the 26th Annual Meeting of the Cognitive Science Society* (2004).
21. Purchase, H. C., Freeman, E., and Hamer, J. An exploration of visual complexity. *Digrammatic representation and inferences* (2012), 200-213.
22. Purchase, H. C., Hamer, J., Jameson, A., and Ryan, O. Investigating objective measures of web page aesthetics and usability. In *12th Australasian user interface conference (AUIC 2011)* (2011), Australian computer society, Inc., 19-28.
23. Reber, R., Schwarz, N., and Winkielman, P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8, 4 (2004), 364-382.
24. Reber, R., Winkielman, P., and Schwarz, N. Effects of perceptual fluency on affective judgments. *Psychological science*, 9, 5 (1998), 45-48.
25. Reinecke, K. and Gajos, K. Z. Quantifying visual preferences around the world. In *the 32nd annual ACM conference on Human factors in computing systems* (2014), ACM.
26. Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., and Liu, J., Gajos, K. Z. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *CHI* (2013), ACM, 2049-2058.
27. Rosenholtz, R., Li, Y., and Nakano, L. Measuring visual clutter. *Journal of vision*, 7, 2 (2007), 1-22.
28. Sereno, S. C. and Rayner, K. Measuring word recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, 7, 11 (2003), 489-493.
29. Sonderegger, A., Zbinden, G., Uebelbacher, A., and Sauer, J. The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. *Ergonomics*, 55 (2012), 713-730.
30. Thielsch, M. T., Blotenberg, I., and Jaron, R. User evaluation of websites: From first impression to recommendation. *Interacting with Computers*, 26, 1 (2013), 89-102.
31. Tinio, P. P. and Leder, H. Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica*, 130, 3 (2009), 241-250.
32. Tractinsky, N. Visual Aesthetics. In Soegaard, Mads and Dam, Rikke Friis, eds., *The Encyclopedia of Human-Computer Interaction, 2nd Ed*. The Interaction-Design.org Foundation, Aarhus, 2013.
33. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International journal of human-computer studies*, 64, 11 (2006), 1071-1083.
34. Tuch, A. N., Bargas-Avila, J. A., and Opwis, K. Symmetry and aesthetics in website design: It'sa man's business. *Computers in Human Behavior*, 26, 6 (2010), 1831-1837.
35. Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. Visual complexity of websites: effects on users' experience, physiology, perfomance, and memory. *International journal of human-computer studies*, 67 (2009), 703-715.
36. Tuch, A. N., Presslaber, E. E., Stocklin, M., Opwis, K., and Bargas-Avila, J. A. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, 70 (2012), 794-811.
37. Van Den Berg, R., Cornelissen, F. W., and Roerdink, J. B. A crowding model of visual clutter. *Journal of Vision*, 9, 4 (2009), 1-11.
38. Van Schaik, P. and Ling, J. The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies*, 67, 1 (2009), 79-89.
39. Wu, O., Hu, W., and Shi, L. Measuring the visual complexities of Web pages. *ACM Transactions on the Web (TWEB)*, 7, 1 (2013), 1-34.
40. Zheng, X. S., Chakraborty, I., Lin, J. J. W., and Rauschenberger, R. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *SIGCHI Conference on Human Factors in Computing Systems* (2009), ACM, 1-10