# Quantification of Interface Visual Complexity

Aliaksei Miniukovich
University of Trento
Via Sommarive, 9 I-38123 Povo, TN, Italy
miniukovich@disi.unitn.it

Antonella De Angeli
University of Trento
Via Sommarive, 9 I-38123 Povo, TN, Italy
deangeli@disi.unitn.it

## ABSTRACT

Designers strive for enjoyable user experience (UX) and put a significant effort into making graphical user interfaces (GUI) both usable and beautiful. Our goal is to minimize their effort: with this purpose in mind, we have been studying automatic metrics of GUI qualities. These metrics could enable designers to iterate their designs more quickly. We started from the psychological findings that people tend to prefer simpler things. We then assumed visual complexity determinants also determine visual aesthetics and outlined eight of them as belonging to three dimensions: information amount (visual clutter and color variability), information organization (symmetry, grid, ease-of-grouping and prototypicality), and information discriminability (contour density and figure-ground contrast). We investigated five determinants (visual clutter, symmetry, contour density, figure-ground contrast and color variability) and proposed six associated automatic metrics. These metrics take screenshots of GUI as input and can thus be applied to any type of GUI. We validated the metrics through a user study: we gathered the ratings of immediate impressions of GUI visual complexity and aesthetics, and correlated them with the output of the metrics. The output explained up to 51% of aesthetics ratings and 50% of complexity ratings. This promising result could be further extended towards the creation of tLight, our automatic GUI evaluation tool.

## Categories and Subject Descriptors

H.5.2 [**Information interfaces and presentation**]: User interfaces – *graphical user interfaces (GUI), evaluation/methodology.*

## General Terms

Measurement, Human Factors.

## Keywords

Visual aesthetics, GUI quality, metric validation, immediate impression.

## 1. INTRODUCTION

Interaction designers nowadays strive to create interfaces that are not only easy-to-use, but also pleasing to the eye. Aesthetics, therefore, has come under investigation in HCI. The relationship between visual complexity and visual aesthetics was well-established in both psychology [25] and HCI [35]. Simpler visual stimuli tend to be perceived as more aesthetically pleasing, probably due to more fluent mental processing of stimuli [40].


*AVI' 14, May 27 - 29 2014, Como, Italy.*
ACM 978-1-4503-2775-6/14/05…$15.00.

Despite the growing interest in aesthetics, little is yet known on how to design interfaces that are indeed aesthetically pleasing. Most design decisions are left to the personal taste of designer, with, often, poor results. The problem becomes more salient when addressing applications that can be designed by non-professional interface designers, such as web sites and mobile apps. To be interpreted correctly, existing aesthetic design guidelines (e.g., [30]) require wide practical experience, and are therefore of little help to non-professional designers. Automatic tools of design aesthetics evaluation, on the other hand, could quickly detect and visualize problematic design aspects.

In this paper, we present an initial study of the development of tLight, an automatic tool for GUI aesthetics evaluation. We started by adapting constructs from psychology, which are known to correlate with perceived complexity and are heavily based on the Gestalt principles. We converted a part of these constructs into a set of automatic metrics, either using existing solutions or suggesting our own algorithms. To test the metrics, a user study was conducted: ten participants rated visual aesthetics and visual complexity of 140 webpage screenshots on a 5-point Likert scale. Participants' complexity ratings accounted for 30% of the aesthetics ratings, thus supporting our initial intent to explain aesthetics with complexity-originated metrics. Our automatic metrics also accounted for a part of complexity-independent variance in aesthetics. The best-fit regression models could explain 50% of variance in perceived visual complexity and 51% of variance in visual aesthetics.

We acknowledge that aesthetics is a complex, multilayered phenomenon (cf. [1]), which consists of many culture-independent and culture-specific facets. We intended to only address culture-independent facets, related to perceived visual complexity, i.e. to the effort of processing stimuli. This effort is thought to closely relate to aesthetic pleasure across all cultures [26] and was at the core of our decisions on experimental design and automatic metrics. The paper is structured as follows: chapter 2 reviews the related work on complexity, aesthetics and GUI evaluation; chapter 3 delves into the exploratory study of GUI visual complexity and aesthetics; chapter 4 discusses the results of the study and highlights future research directions; chapter 5 sums up the results.

## 2. RELATED WORK

People constantly and automatically recognize, recall, memorize, attribute and evaluate, i.e. they process information coming from their surrounding environment. The fluency of this processing largely determines how complex stimuli are perceived and brings forth initial visual aesthetics judgments [25, 40], which partially persist over time [31]. The initial judgments further evolve according to conscious elaboration, which could be a source of more intense aesthetic pleasure relative to processing fluency [1]. However, the mechanisms of how conscious thinking influences aesthetics is much less explored.

The processing of stimuli involves both pre-conscious perception and conscious cognition (cf. [25]), and accordingly, we would expect two types of stimulus complexity in the GUI domain: visual and conceptual. Most studies of GUI quality have implicitly concentrated on the conceptual side, rather than explicitly differentiated between these two types. Ivory et al. [13] proposed and validated a set of page-level metrics of webpage quality, such as the number of words, fonts and links per page, and reading complexity. Although this was not explicitly stated, these metrics substantially resembled the concept of cognitive load [9]. Depending on the website category, they could explain from 11% to 56% of webpage rating variance. Reinecke et al. [27] explained up to 48% of user ratings of webpage visual appeal and 65% of webpage visual complexity. Their automatic complexity metric took into consideration the amount of text per page, the number of text and non-text areas, and images, and colorfulness.

Unlike the studies above, Wu et al. [42] identified low-level visual GUI characteristics (e.g., the number and sizes of visual blocks, and density of text characters), which accounted for 46% of variance in webpage visual quality ratings A part of their measures (e.g., the average values of hue, saturation and value of webpage screenshot), though, could represent the preferences of a particular social group rather than interface complexity, which might lower the generalizability of their results. Purchase et al. [22] also concentrated on the visual side of complexity and operationalized it with the number of image colors (before and after color reduction), the variability in pixel luminance, the ratio of edge pixels to all pixels, and with the sizes of images saved in PNG, GIF and JPEG formats. Although their best-fit model could only explain 25% of complexity ratings, they used pixel-based metrics, without including any semantic, page-element-based metrics: therefore, no dependency on GUI type and cultural context was introduced.

Still, the studies above did not adopt a systematic approach to interface complexity; they often considered only a few of its determinants (e.g., color variability and edge density [22]) or even simply assumed it to correspond to the number of interface elements (e.g., [13]). Psychologists, on the other hand, did approach complexity more systematically, distinguishing between amount of information and organization of information (see also [37]). Still, this might be not enough either. The assessment of visual complexity also depends on the discriminability of information – difficulties not in processing, but in receiving "raw" visual input. Oliva et al. [20] investigated what *visual complexity* meant for the viewers of complex indoor scenes, and found that the viewers substantiated their sorting decisions according to the number of objects, colors and details (the amount of information), clutter, open space, symmetry and organization (the organization of information) and figure-ground contrast (the discriminability of information). Low figure-ground contrast reflects the difficulties in seeing rather than in processing information. Figure 1 shows the classification of visual complexity in three main determinants: amount, organization and discriminability of information.

## 2.1 Amount of Information

The amount of information is determined by scene variation in color, luminance, orientation, motion and other visual features. It was often studied in HCI under the name 'visual clutter' [28, 36]. The variation in color, being the most prominent feature of scene, was also studied separately from visual clutter under the name 'colorfulness' [27].

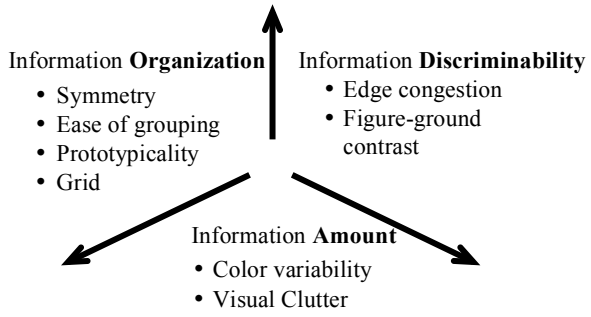


**Figure 1. The classification of visual complexity determinants**

### *2.1.1 Visual clutter*

"*Clutter is the state in which excess items, or their representation or organization, lead to a degradation of performance at some task*" ([28], p.3). Existing measures of visual clutter roughly approximate the number of distinct objects in a scene, which strongly correlates with the search time of target object in a set of distractor objects [3].

Researchers measured visual clutter in several ways. Mack et al. (as cited in [28]) suggested using the ratio of edge pixels to all pixels of an image (the Edge Density measure). Rosenholtz et al. [28] proposed Feature Congestion (the difficulty of introducing a visually salient object) and Subband Entropy (the amount of redundancy encoded in a scene). The algorithm of Subband Entropy resembles the algorithm of the JPEG image compressor, which was also adopted as a simple measure of Visual Complexity [34]. Each of these measures was found to correlate with search performance.

### *2.1.2 Color variability*

Color variability is a very salient feature of scene. In a study [20], the number of colors ranked second among the most important factors of visual complexity (after the number of objects). However, Oliva et al. [20] did not elaborate on the notion of number of colors. Seemingly, participants meant the number of dominant colors (i.e., the colors they could easily identify), which is different from color depth (large variety of eye-unperceivable color shades).

The lack of agreement on what color variability is has resulted in a variety of different measures. Hasler et al. [11] proposed a measure of 'colorfulness', which combined the mean and standard deviation of red-green and blue-yellow color components in the Lab color space. Reinecke et al. [27] clustered image colors into 16 main W3C-defined colors and measured the area they occupied, computed the average of hue, saturation and value over an image, and re-used Hasler et al.'s measure of colorfulness [11]. They could explain 78% of variation in colorfulness ratings. Similarly, Wu et al. [42] used the average values and variation in hue, brightness and saturation, and again, colorfulness [11]. They could explain 46% of variation in webpage visual quality ratings. Lastly, Purchase et al. [22] accounted for almost 25% of variation in the ratings of image visual complexity. They used the number of image colors before color reduction, the number of image colors after color reduction (adopting 3 different color reduction procedures) and standard deviation in pixel luminance.

To summarize, the variety of proposed measures of color variability could be divided into three categories: the number of dominant colors (e.g., Purchase et al.'s [22] image colors after reduction), perceived color depth (e.g., Wu et al.'s [42] variation in hue, brightness and saturation) and idiosyncratic color preferences (e.g., Reinecke et al.'s [27] preferences for W3C-

defined colors). The last category is strongly influenced by acquired tastes [4], and, should therefore be avoided when trying to generalize findings across different demographics.

## 2.2 Organization of Information

Psychology and HCI research provides at least four determinants of visual complexity related to the organization of information category: symmetry, ease of grouping, prototypicality and grid.

### *2.2.1 Symmetry*

One of the Gestalt principles, mirror symmetry – the similarity of an object reflection across a straight axis – was claimed to improve interface design [29]. However, quantifying symmetry might be problematic in HCI. Psychologists mainly studied mirror symmetry of relatively simple objects, such as dot and line patterns, or human faces. In HCI, Bauerly et al. [2] and Tuch et al. [33] varied the amount of global symmetry in webpages, and found a correlation with aesthetics and design preferences. However, they did not measure symmetry; they manually altered webpages to be either symmetrical or asymmetrical. In a different study, Zheng et al. [43] used quadtree decomposition (recursive image partitioning in a tree-like structure of visually homogeneous blocks), and operationalized symmetry as vertical and horizontal mirror reflection of quadtree leafs. Surprisingly, their symmetry measure did not correlate with the ratings on "complicated-simple". After re-using the same algorithm, Reinecke et al. [27] also did not report an influence of symmetry on visual complexity of webpages. Lastly, image-processing scholars adopted a more sophisticated approach: they distinguished local symmetry (i.e., the symmetry over a small area of image) from global symmetry and proposed several algorithms for measuring it [14, 19].

### *2.2.2 Ease of grouping*

Gestalt psychologists discussed in detail what is perceived as a group and how to increase the ease of grouping [38]. Gestalt-based design guidelines [29] instructed designers to place related objects together and make them visually similar. Psychology first offered a strong empirical evidence of benefits of high ease of grouping [32] and described several necessary criteria of grouping, such as common location, motion, color, surface, texture, size, and shape [6, 5]. However, even when the list of criteria is reduced to only two principles – the similarity within a visual group, and the difference between one visual group and other groups – it is still an open question how we should measure them automatically.

### *2.2.3 Prototypicality*

An interface element can be defined as prototypical if an average user sees it as a representative of a class of elements. The profound effect of prototypicality on user perception follows from the work in psychology. One of the Gestalt principles, compliance with past experience or habit [38], states that we see things in the way we are accustomed to see them. Strong deviation from the prototype often results in a negative experience [12], whereas subtle deviations are often appreciated. Pseudohomophones – non-words that sound similar to words – are another example of prototypicality-related source of processing fluency. The word-like composition basis (which participants are vastly familiar with) let participants read pseudohomophones quicker than non-words [39]. However, despite its significant role in user perception, we found no description of automatic prototypicality measurements. The work on large-scale design mining [15] is a potential step in this direction. It allows studying existing design practices and might serve as a basis for future prototypicality measurements.

### *2.2.4 Grid*

Most graphical user interfaces are based on a grid, i.e., leverage regular repetition of similar structural elements. Regular repetition might be as equally important as symmetry for figural goodness and visual simplicity [21]. Only a few studies, though, exploited it in studying interface structure. Reinecke et al. [27] used quadtree-based symmetry, balance and equilibrium, but not repetition, and reported no influence of these factors on visual complexity. Wu et al. [42] operationalized layout complexity as the number of leafs and number of levels of webpage visual block tree, but still did not exploit repetition. Lastly, Harper et al. [8] placed top-left corners of webpage elements in an overlay grid, and computed the average number of corners in a grid cell and variation of corner amounts across the grid. They found a Spearman correlation of $r = 0.95$ for their manual and computational rankings of 20 webpages. However, the notion of repetition was not used. We believe a more sophisticated metric of grid might be needed, which, for instance, accounts for alignment, regularity and uniform separation (see [10] for a short description of these factors).

## 2.3 Discriminability of Information

The human visual system has natural limits, e.g., the minimal perceivable luminance difference between two areas or minimal perceivable dot size. Such limitations influence the discriminability of information and might cause higher visual complexity even if the amount and organization of information stay the same. Edge congestion and low figure-ground contrast are two aspects of information discriminability.

### *2.3.1 Edge congestion*

Discriminating and tracing a line in a line congestion situation can be problematic. This issue often emerges in the domain of large-graph visualization. Wong et al. [41] stated the problem: "*the density of edges is so great that they obscure nodes, individual edges and even visual information beneath the graph*". They also proposed an interactive solution to edge congestion in graphs – the edges were bent away from users' point of attention without changing the number of nodes or edges. A more sophisticated discussion of edge congestion comes from the research on crowding and is grounded on the notion of critical spacing – the distance between objects at which object perception starts to degrade [16]. For example, the crowding model of visual clutter [36] uses eccentricity-based critical spacing to account for information loss in peripheral visual field. However, the accompanying algorithm accounts simultaneously for both visual clutter and edge congestion, and might need reconfiguration to account for edge congestion only.

### *2.3.2 Figure-ground contrast*

Psychologists often use luminance or color contrast to manipulate perceptual fluency. For example, Reber et al. [23] showed participants phrases in green or red on a white background (high contrast condition), and in yellow or light-blue color on a white background (low contrast condition). The high-

contrast phrases were judged as true facts significantly above the chance level, whereas the low-contrast phrases were not. The authors attributed it to the difference in reading difficulty, and thus, processing fluency. Similarly, Reber et al. [24] showed participants 70% black (high contrast) and 30% black (low contrast) words on a white background. In the high contrast scenario, the participants were significantly faster at detecting and recognizing words. Hall et al. [7] explored text readability of web pages, and found white-black text-background combinations to be more readable than light- and dark-blue, or cyan-black combinations. However, the studies above did not measure contrast automatically.

# 3. EXPLORATORY STUDY

We undertook an exploratory study of GUI visual complexity and aesthetics. First, we selected webpages and took screenshots of them. Then, we collected user ratings of webpage visual complexity and aesthetics. We deliberately targeted users' immediate impressions, which are heavily influenced by low-level, pre-semantic qualities of UIs and only depend on cultural preferences to a limited extent (cf. [17]). Finally, we computed six automatic screenshot-based metrics and compared them against user ratings. Visual complexity determinants (Figure 1) that are not explored in the paper were left for future work.

## 3.1 Data collection

We analyzed screenshots of 140 webpages. We only considered homepages of websites in English with little or no animation and dynamic effects. A total of 115 of them were found on four public showcases[1] featuring beautiful websites from four categories: a) coffee, b) chocolate bars and shops, c) online retailers and d) design agencies. This sample was obviously biased towards the beautiful design. To counterbalance the emphasis on beauty, we added 25 more websites from similar categories, which we considered unappealing. We took full-length screenshots of web-pages in the PNG format (truecolor, 24-bit per pixel) and cropped them to fit the screen (the top part only; 1280×800 pixels).

Ten graduate students (mean age = 27.2 years, SD = 2.04; 7 – male; all fluent in English) of the local university participated in the study. The study was conducted individually in a separate room with a laptop. A researcher was always present in the room. Participants were instructed to rate webpages on a 1-5 Likert scale according to "how simple/complex the webpages were" and "how unattractive/attractive the webpages were" (cf. [18, 33, 34, 35]), without any further explanations of the constructs (cf. [20]). After participants signed the consent forms and adjusted their seat height and screen angle, they rated 140 (cf. [17, 18, 31, 35]) screenshots (5 screenshots were used for training). The entire procedure took ~25 min; no complaints about fatigue were reported. The experiment was run at the 1280×800 pixel resolution and 60 Hz refresh rate. We coded experimental procedures using PsychoPy[2].

Each participant saw and rated each screenshot only once, one after another. First, we asked participants to focus the sight on a red fixation cross on gray background shown for 1-1.5 seconds. Second, a screenshot was flashed for 50msec (cf. [18, 17, 35]). Third, a black-white noise mask was flashed for 50msec (cf. [35]) to cancel out extended visual perception. Fourth, participants rated visual complexity and aesthetics of webpages with no time constraints. Ratings were set using the 1-5 keyboard buttons. Both questions addressing complexity and aesthetics were shown on the screen simultaneously. Screenshot presentation was randomized. The very short presentation span (50ms) was intended to ensure we measured complexity at a perceptual level. Longer time spans would allow cognitive elaboration and cause higher individual variability in judgments (cf. [31]).

## 3.2 Automatic metrics

Out of eight identified determinants of visual complexity (Figure 1), we investigated five: visual clutter, color variability, symmetry, edge congestion and figure-ground contrast. The investigation resulted in six metrics (color variability had two distinct aspects: number of dominant colors and color depth). Visual clutter and color variability metrics consisted of multiple measures; symmetry, congestion and contrast consisted of a single measure each. The metrics were implemented in Matlab and took screenshots of GUIs as input, i.e. they could be applied to any graphical interfaces, not only webpages.

### *3.2.1 Amount of information*

From the amount of information, we modeled both visual clutter and color variability. Our visual clutter metric combined four measures from the literature. First, we took the ratio of edge pixels to all pixels (CL1, see Edge Density in [28]). Edges were detected with the Canny method; low and high thresholds were set to 0.11 and 0.27 (cf. [28]); the standard deviation of a Gaussian for pre-detection smoothing was $\sqrt{2}$. Then, we calculated Subband Entropy (CL2) and Feature Congestion (CL3) with authors' settings[3] (adapted for geographical map analysis, [28]). Then, we took file sizes of screenshots saved in JPEG (CL4, the JPEG quality setting set at 70, cf. [34]). Finally, we conducted the maximum-likelihood factor analysis with Varimax rotation on the measures (CL1-CL4). All measures loaded on a single factor (Table 1A) and were combined in a single metric of visual clutter. To generate combined scores, we used Thomson's method, which maximizes score determinacy.

**Table 1. Factor loadings of A) clutter (cumulative var. g = .82) and B) color variability (cumulative var. g = .63)**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| A | Factors | CL1 | | CL2 | | CL3 | | CL4 | | |
| | Clutter | .91 | | .86 | | .91 | | .93 | | |
| B | | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 |
| | Color depth | .73 | .89 | .89 | .92 | .89 | .20 | -.11 | | .40 |
| | Dominant colors | 0,22 | | | | | .97 | .60 | .49 | .33 |

Literature describes many measures of color variability [27, 42, 22, 11]. We did not consider those that might involve subjective color preferences (e.g., [27]) and only took the measures that could account for the number of dominant colors or perceived color depth. A part of these measures was based on the number of colors before color reduction and had very skewed distribution (e.g., a 1280*800 image almost always span over all 256 available hues and have $\chi^2$-like distribution). Since data normality is required for regression and factor analyses, we excluded these measures from further analysis. The final set of measures is shown in Table 2. To investigate data dimensionality, we

[1] http://www.smashingmagazine.com/
[2] http://www.psychopy.org/
[3] We used authors' implementation of Subband Entropy and Feature Congestion [28] cfr. http://dspace.mit.edu/handle/1721.1/37593

**Table 2. Implemented measures of color variability.**

| # | Measure name | Measure description |
|---|---|---|
| C1-C3 | Hue, saturation and value after color reduction. | The number of distinct values of Hue, Saturation and Value. Images were converted to the HSV color space. Color variability was reduced: only values covering more than 0.1% of image were counted. |
| C4 | RGB colors after color reduction. | The number of distinct RGB values. Color variability was reduced: only colors covering more than five pixels were counted. |
| C5 | PNG file sizes. | The file sizes of screenshots, saved in PNG format (24 bits per color). |
| C6 | Static clusters of RGB colors after color reduction. | The number of static clusters of RGB values ($32^3$ combinations per cluster, 512 clusters maximum). Color variability was reduced: only colors covering more than five pixels were counted. |
| C7 | Dynamic clusters of RGB colors after color reduction. | The number of dynamic clusters of RGB values (Figure 2). After color variability is reduced (only colors covering more than five pixels), a between-color difference is considered. If the distance in all color components is less than three, two colors are united in the same cluster. Uniting continues recursively till all used colors are assigned to a cluster. |
| C8 | Colorfulness [11] | The measure of colorfulness from [11]. It combines standard deviations and means of pixel values taken in the Lab color space. |
| C9 | Luminance SD | The standard deviation of pixel luminance. |

conducted maximum-likelihood factor analysis with Varimax rotation. All measures but C8 (colorfulness) and C9 (luminance SD) loaded on either of two main factors (Table 1B). C8 and C9 were excluded from further analysis. The measures C1-C7 were combined in Bartlett's scores of color depth (Factor 1) and dominant colors (Factor 2). We expected the orthogonality of color depth and dominant color factors, and therefore used Bartlett's score generation method rather than Thomson's method. Bartlett's method maximizes the independence of orthogonal factor variances.

### *3.2.2 Organization of information*

As regards the organization of information, we modeled mirror symmetry. The existing methods for symmetry detection [19] did not give satisfactory results on our data. They often detected many false symmetry axes indistinguishably from true symmetry axes[4]. In HCI, the presence of screen frame calls for accounting for horizontal and vertical symmetry only (cf. [33]). We proposed an algorithm, which detected the mirror symmetry along the central vertical axis. First, our algorithm detected image contours based on the Canny edge detection algorithm with the low and high thresholds set to 0.11 and 0.27 (cf. [28]) and the standard deviation of a Gaussian for pre-detection smoothing set to 5 (high Gaussian SD allows detecting contours of text blocks rather than of individual characters). Then, the algorithm took only the vertical component of detected contours: horizontal lines across the central axis always give symmetrical key points, see Figure 2. We reduced the number of contour pixels (by taking a contour pixel and dismissing others in the 3-pixel radius) and took them as key points. Further, for each key point, the algorithm looked for a match in the 4-pixel-radius area across the central axis. Then, we took the ratio of matches ($k_{sym}$) to all key points ($k_{all}$) and normalized this ratio by the probability of key point match due to chance (more key points in the same area mean higher probability of match due to chance). The probability of incidental match depends on the number of all key points, the size of match-search area ($S_a$, which was constant for all screenshots, 4-pixel-radius area) and size of screenshots ($S_s$, which was also constant across all screenshots). The normalized ratio ($Sym_{norm}$) was the metric of symmetry we used:

$$Sym_{norm} = \frac{k_{sym}}{k_{all}} * \left( \frac{(k_{all} - 1) * S_a}{S_s} \right)^{-1}$$

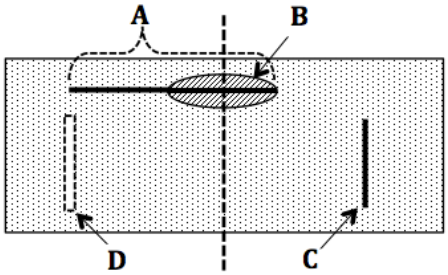


**Figure 2. An asymmetrical horizontal line (A) can give symmetrical key points (B); whereas a vertical line (C) cannot, unless it matches another line (D) on the other side.**

### *3.2.3 Discriminability of information*

We operationalized edge congestion and figure-ground contrast. Edge congestion relies on the notion of critical spacing – minimal distance between objects at which the user starts having difficulties differentiating the objects. We tried out 8-, 12- and 20-pixel thresholds, which all gave similar results. Hence, we chose the 20-pixel threshold. The algorithm consisted of two main steps: detecting edges and detecting edges in close proximity. Any edge detection algorithm without pre-detection smoothing could be used (e.g., the Sobel or Prewitt algorithms); we used simple value difference of more than 50 between adjacent pixels across all three (red, green and blue) color components. The edges by different color components were combined using disjunction. In the second step, if at least two pixels of two different edges occurred in the same 20-pixel vicinity, they were marked as congested. The marking was done in both horizontal and vertical directions. Finally, we took the ratio of 'congested' pixels ($p_c$) to all edge pixels ($p_a$) as the metric of edge congestion: $cong = p_c/p_a$. Edge congestion does not require chance normalization as the symmetry metric above, despite they both leverage edge detection. Whatever the reason is two edges are too close, they still impede user perception fluency.

Figure-ground contrast describes the difference in color or luminance between two adjacent areas. This difference forms an

[4] We used authors' implementation of the algorithm [19] cfr. http://www.nada.kth.se/~gareth/homepage/local_site/code.htm

edge; the magnitude of the difference defines the strength of the edge. To detect edges, we used the Canny edge detection algorithm, which requires two input thresholds in the range of 0 to 1: low and high. Lower input thresholds allow detecting more edges (i.e., both weak and strong edges); higher thresholds allow detecting fewer edges (i.e., only strong edges). We tried out a variety of different threshold settings and rejected large values, which afforded detecting too few edges. Our low threshold was always 40% of the high threshold. The high threshold varied from 0.1 to 0.7 with a step of 0.1, which gave us seven levels ($l$) of edge strength ($E_l$). We counted edge pixels for each level and computed the difference between successive levels. Then, we weighted the differences and summed them up. The weakest edges received the highest weight, since they contribute the most to the difficulty of visual differentiation. Lastly, the sum was normalized by the number of all edge pixels (i.e., the edges detected with the high threshold 0.1 minus the edges detected with the high threshold 0.7). The normalized sum ($E_{norm}$) was the metric of figure-ground contrast:

$$E_{norm} = \frac{\sum_{l=1}^{6}(E_l - E_{l+1})}{E_1 - E_6} * (1 - \frac{(l-1)}{6})$$

## 3.3 Results

We gave participants no description of complexity or aesthetics. However, the average scores of interclass correlation coefficients were satisfactory for both complexity (ICC = .69; 95% conf. interval 0.609 < ICC < 0.762) and aesthetics (ICC = .81; 95% conf. interval 0.755 < ICC < 0.851), indicating interrater reliability. We calculated rating means for each webpage and used them in a further analysis as webpage complexity and aesthetics scores.

The means of complexity and aesthetics scores were $m_c = 2.69$ ($SD_c = .53$) and $m_a = 2.98$ ($SD_a = .66$); aesthetics scores were slightly negatively skewed ($sk_a = -.54$), reflecting biased selection criteria favoring beautiful web-pages. As expected, complexity scores negatively correlated with aesthetics scores (r = -.55; p-value < .001) and could explain 30% of aesthetics score variance ($R^2 = .30$; $b_{const} = 4.83$; $b_{comp} = -.69$; $SE_{const} = .24$; $SE_{comp} = .09$; p-value < .001).

We calculated correlations of complexity and aesthetics scores with our six automatic metrics (Table 3). All six metrics were intended to account for different aspects of perceived visual complexity, and only one metric (color depth) did not. Four metrics (visual clutter, dominant colors, contrast and edge congestion) correlated with both aesthetics and complexity, suggesting a mediation effect of complexity on aesthetics (i.e., the metrics influenced aesthetics indirectly: they influenced complexity, which, in turn, influenced aesthetics). The mediation effect was indeed observed, see Table 4. When taking complexity into account, the visual clutter influence on aesthetics was no longer significant, thus suggesting full mediation. The influence of dominant colors, contrast and congestion on aesthetics was still significant but smaller, suggesting partial mediation. The best-fit model perceived of visual complexity explained 50% of complexity scores, see Table 5a; the best-fit model of visual aesthetics explained 51% of aesthetics scores, see Table 5b. Both models included three independent variables and shared two of them: number of dominant colors and edge congestion. Visual clutter was solely relevant to complexity; color depth was solely relevant to aesthetics. Adding visual complexity scores as an independent variable to the aesthetics model gave only 6% increase in the explained variance, see Table 5c.

**Table 3. Pearson correlation coefficients**

| | Aesthetics scores | Complexity scores |
|---|---|---|
| Dominant colors | -.47*** | .61*** |
| Color depth | .45*** | .01 |
| Visual clutter | -.24** | .56*** |
| Contrast | .48*** | -.32*** |
| Edge congestion | -.53*** | .46*** |
| Symmetry | -.01 | -.24** |

** p < .01; *** p < .001.

**Table 4. Mediator analysis (complexity - mediator; aesthetics - DV): β-scores of linear models before and after considering the mediator, and magnitude of indirect effect (Sobel test).**

| Independent variables | | β | Indirect effect[1] |
|---|---|---|---|
| Visual Clutter | before | -.24** | full mediation |
| | after | .10 | |
| Dominant colors | before | -.47*** | -.26 |
| | after | -.21* | |
| Figure-ground Contrast | before | .48*** | .14 |
| | after | .34*** | |
| Edge Congestion | before | -.53*** | -.18 |
| | after | -.35*** | |

* p < .05; ** p < .01; *** p < .001; [1] based on z-scores.

**Table 5. Regression models of a) perceived visual complexity ($R^2$ = .50), b) visual aesthetics ($R^2$ = .51) and c) visual aesthetics with complexity included ($R^2$ = .57)**

| | Ind. variable | Predictor | β | t-value |
|---|---|---|---|---|
| A | Visual complexity | (Intercept) | | 9.85 |
| | | Dominant colors | .39*** | 5.46 |
| | | Visual clutter | .29*** | 4.06 |
| | | Edge congestion | .23** | 3.43 |
| B | Visual aesthetics | (Intercept) | | 15.22 |
| | | Color depth | .38*** | 6.27 |
| | | Dominant colors | -.35*** | -4.20 |
| | | Edge congestion | -.32** | -4.08 |
| C | Visual aesthetics | (Intercept) | | 15.16 |
| | | Color depth | .41*** | 6.41 |
| | | Dominant colors | -.18* | -2.31 |
| | | Edge congestion | -.22** | 3.14 |
| | | Visual Complexity | -.35*** | -4.20 |

* p < .05; ** p < .01; *** p < .001.

# 4. DISCUSSION

In this paper, we presented an exploratory study of perceived visual complexity and aesthetics and their relationship with five GUI complexity determinants. We operationalized three determinants (contrast, edge congestion and symmetry) with single-item metrics and the other two (visual clutter and color variability) with multi-item metrics. The employed measures of visual clutter had very high loadings (Table 1A) on a single factor, allowing us to combine them in a single complex metric of clutter. A similar analysis of color variability (Table 1B) revealed two prominent factors: the number of dominant colors and perceived color depth. Dominant colors positively correlated with perceived

complexity and negatively with aesthetics; perceived color depth positively correlated with aesthetics and did not correlate with perceived complexity. We emphasize the distinction between dominant colors and color depth: if they are combined in a single measure of 'colorfulness', they explain little of complexity and aesthetics scores [27].

The results of our user study were in line with similar work [42, 27]: GUI visual complexity indeed affected immediate aesthetics impression. Yet, complexity scores alone explained only 30% of aesthetics score variance, whereas our automatic metrics explained up to 51%. Four of the metrics (perceived color depth, dominant colors, figure-ground contrast and edge congestion) correlated with aesthetics scores and were not fully mediated by complexity. This could imply they captured GUI qualities unrelated to complexity but relevant to aesthetics. Further studies are needed to investigate the origin and generalizability of this non-complexity-based aesthetics.

Nonetheless, the mediation effect of complexity was prominent (Table 4), meaning the same GUI transformation could simultaneously decrease complexity and increase aesthetics. Contrast, edge congestion and dominant colors had a profound effect on both complexity and aesthetics scores (Table 3) and were only partially mediated by complexity. We emphasize them as prominent sources of GUI improvement in both complexity and aesthetics aspects. The effect of visual clutter on aesthetics was smaller and fully mediated by complexity. Still, this converged with the prior use of automatically measured clutter instead of subjective complexity [35].

Considering individual correlations, our symmetry metric only moderately correlated with complexity scores and did not correlate with aesthetics scores. This was in line with previous work (e.g., [27]) and suggested that participants might not consider symmetry while judging aesthetics or that other effects (e.g., demographic, see [33]) could be present. Our contrast metric correlated in the unexpected direction with complexity and aesthetics scores. Participants seemingly preferred low-contrast contours to high-contrast contours and attributed them rather to simplicity than complexity. Further studies are needed to test the nature of this effect. Edge congestion, visual clutter and the number of dominant colors correlated positively with complexity scores and negatively with aesthetics scores, as expected.

Remarkably, our results only partially supported previous use of image file sizes as a measure of visual complexity [34, 22]. Whereas the JPEG file sizes described visual clutter (cf. [28]), and through that, visual complexity, the PNG file sizes described color depth, the measure unrelated to visual complexity, Table 3. This suggested that PNG files sizes are ineffective in describing visual complexity.

The present work is a step towards tLight, a system of automatic GUI evaluation. However, the development of a fully functional system requires further effort. First, future studies will need to extend our results to the other interface types (e.g., mobile app interfaces), which might be influenced in a different way by the GUI qualities we explored. Second, future studies will need to include three more determinants of visual complexity (Figure 1): GUI prototypicality, grid and easiness of grouping. In addition, the proposed dimensionality of visual complexity (Figure 1) needs to be tested. Lastly, future studies might explore if leveraging GUI-level features (e.g., the number of buttons or font sizes) rather than pixel-level features can explain additional variance in user preferences.

# 5. CONCLUSION

In this paper, we outlined eight low-level determinants of GUI complexity (Figure 1). Then, we investigated five determinants and proposed six associated pixel-based metrics (one determinant, colorfulness, was measured by two metrics). We attempted to maximize the use of existing measures and algorithms in our metrics. However, the development of metrics of contrast, edge congestion and symmetry also required the development of our own algorithms. We tested the metrics on 140 webpage screenshots, and explained 50% of variation in the user ratings of visual complexity and 51% of variation in the user ratings of visual aesthetics. The practical application of metrics is twofold: general testing if a GUI is beautiful and visually simple, and finding dimensions a design performs particularly badly on.

# 6. REFERENCES


[1] Armstrong, T. and Detweiler-Bedel, B. 2008. Beauty as an emotion: the exhilarating prospect of mastering a challenging world. *Review of General Psychology*, 12, 4, 305-329.

[2] Bauerly, M. and Liu, Y. 2006. Effects of symmetry and number of compositional elements on interface and design aesthetics. In *the Human Factors and Ergonomics Society Annual Meeting*, 304-308.

[3] Bravo, M. J. and Farid, H. 2004. Search for a category target in clutter. *Journal of Perception*, 33, 6, 643-652.

[4] Cyr, D., Head, M., and Larios, H. 2010. Colour appeal in website design within and across cultures: A multi-method evaluation. *International Journal of Human-Computer Studies*, 68, 1, 1-21.

[5] Duncan, J. and Humphreys, G. W. 1989. Visual search and stimulus similarity. *Psychological Review*, 96, 3, 433-458.

[6] Duncan, J. 1984. Selective attention and the organization of visual information. *Journal of Experimental Psychology*, 113, 4, 501-517.

[7] Hall, R. H. and Hanna, P. 2004. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioral intention. *Behaviour & Information Technology*, 23, 3, 183-195.

[8] Harper, S., Jay, C., Michailidou, E., and Quan, H. 2013. Analysing the visual complexity of web pages using document structure. *Behaviors & Information Technology*, 32, 5, 491-502.

[9] Harper, S., Michailidou, E., and Stevens, R. 2009. Toward a definition of visual complexity as an implicit measure of cognitive load. *ACM Transactions on Applied Perception*, 6, 2 (February 2009).

[10] Harrington, S. J., Naveda, J. F., Jones, R. P., Roetling, P., and Thakkar, N. 2004. Aesthetic measures for automated document layout. *the 2004 ACM symposium on Document engineering* (October 2004), 109-111.

[11] Hasler, D. and Suesstrunk, S. 2003. Measuring Colourfulness in Natural Images. In *SPIE/IS&T Human Vision and Electronic Imaging*, 87-95.

[12] Hekkert, P., Snelders, D., and Wieringen, P. C. 2003. 'Most advanced, yet acceptable': typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology*, 94, 1, 111-124.

[13] Ivory, M. Y., Sinha, R. R., and Hearst, M. A. 2001. Empirically validated web page design metrics. In

*Proceedings of the SIGCHI conference on Human factors in computing systems* (Seattle 2001), ACM, 53-60.

[14] Kootstra, G., Bart de B., and Schomaker, L. R. B. 2011 Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 223-240.

[15] Kumar, R., Satyanarayan, A., Torres, C., Lim, M., Ahmad, S., Klemmer, S. R., and Talton, J. O. 2013. Webzeitgeist: design mining the web. In *the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 3083-3092.

[16] Levi, D. M. 2008. Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48, 5, 635-654.

[17] Lindgaard, G., Dudek, C., Sen, D,, Sumegi, L., and Noonan, P. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *Transactions on computer-human interaction*, 18, 1, 1-30.

[18] Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. 2006. Attention web designers: You have 50 milliseconds to make a good first impression! *Behavior & Information Technology*, 25, 2, 115-126.

[19] Loy, G. and Eklundh, J. O. 2006. Detecting symmetry and symmetric constellations of features. In *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg.

[20] Oliva, A., Mack, M. L., Shrestha, M., and Peeper, A. 2004. Identifying the perceptual dimensions of visual complexity of scenes. *The 26th Annual Meeting of the Cognitive Science Society* (August 2004).

[21] Pothos, E. M. and Ward, R. 2000. Symmetry, repetition, and figural goodness: An investigation of the weight of evidence theory. *Cognition*, 75, 3, B65-B78.

[22] Purchase, H. C., Freeman, E., and Hamer, J. 2012 An exploration of visual complexity. *Diagrammatic Representation and Inferences*, 200-213.

[23] Reber, R., Winkielman, P., and Schwarz, N. 1999. Effects of perceptual fluency on affective judgments. *Psychological Science*, 9, 1, 45-48.

[24] Reber, R., Wurtz, P., and Zimmermann, T. D. 2003. Exploring “fringe” consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition*, 13, 1, 47-60.

[25] Reber, R., Schwarz, N., and Winkielman, P. 2004. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8, 4, 364-382.

[26] Reber, R. 2012. Processing fluency, aesthetic pleasure, and culturally shared taste. In *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford University Press, Oxford.

[27] Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., and Gajos, K. Z. 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *CHI* (Paris 2013), ACM, 2049-2058.

[28] Rosenholtz, R., Li, Y., and Nakano, L. 2007. Measuring visual clutter. *Journal of Vision*, 7, 2, 1-22.

[29] Smith-Gratto, K. and Fisher, M. M. 1998-99 Gestalt theory: a foundation for instructional screen design. *Journal of Educational Technology Systems*, 27, 4, 361-372.

[30] Sutcliffe, A. 2009. Designing for user engagement: Aesthetic and attractive user interfaces. *Synthesis lectures on human-centered informatics*, 2, 1, 1-55.

[31] Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. 2006. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International Journal of Human-Computer Studies*, 64, 11, 1071-1083.

[32] Treisman, A. 1982. Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance*, 8, 2, 194-214.

[33] Tuch, A. N., Bargas-Avila, J. A., and Opwis, K. 2010. Symmetry and aesthetics in website design: It’sa man’s business. *Computers in Human Behavior*, 26, 6, 1831-1837.

[34] Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. 2009. Visual complexity of websites: effects on users' experience, physiology, performance, and memory. *International Journal of Human-Computer Studies*, 67, 703-715.

[35] Tuch, A. N., Presslaber, E. E., Stocklin, M., Opwis, K., and Bargas-Avila, J. A. 2012. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *International Journal of Human-Computer Studies*, 70, 794-811.

[36] van den Berg, R., Cornelissen, F. W., and Roerdink, J. B. 2009. A crowding model of visual clutter. *Journal of Vision*, 9, 4, 1-11.

[37] van der Helm, Peter A. 2000. Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, 126, 5, 770-800.

[38] Wertheimer, M. 1938. Laws of Organization in Perceptual Forms (partial translation). In *A Sourcebook of Gestalt Psychology*. Harcourt Brace.

[39] Whittlesea, B. W. and Williams, L. D. 1998. Why do strangers feel familiar, but friends don't? A discrepancy-attribution account of feelings of familiarity. *Acta Psychologica*, 98, 2, 141-165.

[40] Winkielman, P., Schwarz, N., Fazendeiro, T. A., and Reber, R. 2002. The hedonic marking of processing fluency: Implications for evaluative judgment. In *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*. Psychology Press.

[41] Wong, N., Carpendale, S., and Greenberg, S. EdgeLens: 2003. An Interactive Method for Managing Edge Congestion in Graphs. *IEEE Symposium on Information Visualization* (October 19-21, 2003), 51-58.

[42] Wu, O., Chen, Y., Li, B., and Hu, W. 2011. Evaluating the visual quality of web pages using a computational aesthetic approach. In *the fourth ACM international conference on Web search and data mining*, ACM, 337-346.

[43] Zheng, X. S., Chakraborty, I., Lin, J. J. W., and Rauschenberger, R. 2009. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1-10.