Predicting Users' First Impressions of Website Aesthetics With a Quantification of Perceived Visual Complexity and Colorfulness

Katharina Reinecke[1], Tom Yeh[2], Luke Miratrix[1], Rahmatri Mardiko[3], Yuechen Zhao[1], Jenny Liu[1], Krzysztof Z. Gajos[1]

[1]Harvard University
Cambridge, MA, USA
{reinecke,kgajos}@seas.harvard.edu
{yuechenzhao,jennyliu}@college.harvard.edu

[2]University of Colorado at Boulder
Boulder, CO, USA
tom.yeh@colorado.edu

[3]University of Maryland
College Park, MD, USA
mardiko@cs.umd.edu

ABSTRACT

Users make lasting judgments about a website's appeal within a split second of seeing it for the first time. This first impression is influential enough to later affect their opinions of a site's usability and trustworthiness. In this paper, we demonstrate a means to predict the initial impression of aesthetics based on perceptual models of a website's colorfulness and visual complexity. In an online study, we collected ratings of colorfulness, visual complexity, and visual appeal of a set of 450 websites from 548 volunteers. Based on these data, we developed computational models that accurately measure the perceived visual complexity and colorfulness of website screenshots. In combination with demographic variables such as a user's education level and age, these models explain approximately half of the variance in the ratings of aesthetic appeal given after viewing a website for 500ms only.

Author Keywords

Website Aesthetics; First Impression; Colorfulness; Complexity; Modeling; Prediction; Perception

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

INTRODUCTION

It might not be surprising that website aesthetics are a decisive factor for engaging users online. What is surprising, however, is the speed at which we decide whether we like a site or not: According to a series of studies in recent years, users establish a lasting opinion about a website's appeal within a split second of seeing it for the first time [23, 32, 35], and before consciously noticing any details of the design or content. Moreover, if we perceive a website as unappealing, we are less likely to trust it, and more likely to leave it in favor of others [11, 22].

This has enormous economic implications for online consumer-vendor relationships, mirrored in the growing number of research efforts investigating website aesthetics [38, 22, 35]. Despite this interest, no robust methods currently exist to predict a design's aesthetic appeal, leaving designers with qualitative guidelines that are often at the level of anecdotal examples.

Our long-term goal is to develop quantitative models for predicting users' first impressions of aesthetic qualities. Although it is not yet known what exactly influences this first impression of appeal, colorfulness and visual complexity have been repeatedly found to be the most noticeable design characteristics at first sight [27, 38, 34, 8]. Building on these findings, this paper introduces perceptual models of visual complexity and colorfulness in websites, which we then use to estimate users' perception of appeal. Our approach is based on the assumption that this first impression can be adequately captured with the help of a low-level image analysis of static website screenshots.

We make the following contributions:

- We conducted three online experiments to collect colorfulness, complexity, and overall visual appeal ratings from 548 volunteers. Utilizing these data, we developed models that accurately predict perceived visual complexity and perceived colorfulness in websites based on computational image statistics.
- We demonstrate that the predictions of our colorfulness and complexity models can account for nearly half of the variance in the observed ratings of visual appeal. Our results verify the importance of visual complexity for users' first impressions of appeal at first sight. They also demonstrate that colorfulness plays a role, but is not nearly as important as the overall visual complexity of a site.
- Our results also show that the impact of complexity and colorfulness on users' first impressions is not universal, but that age, gender, and education level all influence their first impressions of websites.

CHI 2013, April 27–May 2, 2013, Paris, France.

In the following, we will first review the general related work on aesthetics, before detailing previous work on two of the main influences of appeal—colorfulness and visual complexity. Based on this previous work, we develop a set of image features, and validate them in our first two experiments. The main section then investigates users' first impression of appeal, and the ability of our colorfulness and visual complexity models to predict these first impressions. We close with a discussion, limitations, and future work.

RELATED WORK

For a long time, designers and researchers alike assumed that usability was the main reason for active involvement with a website. It is now known that our initial aesthetic response to products—the spontaneous emotional reaction we have based on our visual preferences—heavily influences whether we later perceive those products as usable [31]. This aesthetic response seems to not only precede judgements about websites, it also influences them in an interaction that has been referred to as the "halo effect": Websites that are perceived as beautiful are also perceived as usable [23, 14, 28, 22] and trustworthy [22]. Since users make reliable judgments within the first 50 to 500ms [23, 32], we believe that it must be possible to employ low-level image statistics of static website screenshots to predict whether a user will like the site.

Prior work in this direction mostly focused on extracting aesthetic features in natural images and evaluating their usefulness with ratings of appeal [9, 10]. Desnoyer and Wettergreen [10], for instance, showed that computational metrics for a photograph's spatial structure, complexity, color features, and contrast can approximate the ratings by an online crowd. However, due to the compositional differences between websites and photographs, such features do not necessarily correspond to our aesthetic perception of websites.

Ivory and Sinha [16] have demonstrated that expert judgements by more than 100 Webby Award judges on the visual appeal of websites can be approximated using 11 predictor metrics (e.g., word and link count, the size of a page in bytes, or the number of different fonts). They achieved 65% accuracy in predicting whether a webpage will be rated very high or very low by the Webby Award judges. The authors acknowledge that it is unclear whether the metrics resemble what judges actually base their opinions on. More importantly, the metrics do not necessarily represent a human's perception of a site, since they are based on information derived from HTML code and cascading style sheets.

Analyzing a screenshot of a website is arguably a better representation of what a user sees and responds to. Zheng et al. [38] used low-level image statistics to analyze the perceived layout structure of a website by calculating symmetry, balance (the optical weight received by either side of the vertical or horizontal image axis), and equilibrium (the centering of interface elements around the image's midpoint). With the goal to evaluate whether users' rapid judgements of a website correlate with low-level image statistics, they asked 22 participants to rate 30 webpages after a brief exposure time of 150ms. While some of their metrics were demonstrated to correlate with the subjective ratings of participants on four dimensions of appeal, it is unclear whether they correspond to what participants perceived. Nevertheless, Zheng et al's work supports previous assumptions that users form an opinion about a website's appeal pre-attentively, before the brain has time to consciously evaluate the incoming stimulus [23].

This paper extends these works by providing the first perceptual models of visual complexity and colorfulness in websites, and demonstrating that these two models alone can already explain much of the variance in users' immediate aesthetic judgements of websites.

QUANTIFYING COLORFULNESS AND COMPLEXITY

A review of related literature identified two promising factors for predicting perceived visual appeal: visual complexity and colorfulness. The following paragraphs therefore focus on these two website characteristics, and subsequently describe the image metrics that we implemented to quantify them.

Colorfulness

One of the most notable features to invoke an emotional reaction is color [21, 5, 25, 8]. Color has been shown to influence perceived trustworthiness [18, 8], users' loyalty [6, 8], and purchase intention [13]. We perceive colors in their entirety, noting various attributes, such as hue (the purity of a color with regards to the primary colors red, blue, and yellow), saturation (the intensity of a color), and luminance (the visually perceived brightness, with yellow having a high luminance, and blue having the lowest luminance value on the color wheel). These features are best described with the perceptually-based HSV model, which comprises a color's *hue*, *saturation*, and *value* (the latter is often used interchangeably with luminance or brightness).

The perceived *colorfulness* is highly dependent on the distribution of colors in an image, and the composition of neighboring colors. For example, two adjacent complimentary colors (i.e., colors that cancel each other's hue) will appear brighter, and can potentially increase the overall perceived brightness. Along with the number and variety of colors in an image, a high brightness can increase our perception of colorfulness.

For natural images, Yendrikhovskij et al. [37] computed the perceived colorfulness as a sum of the average saturation value and its standard deviation across an image. The authors were able to show an extremely high correlation ($r = .91$) between their computational image metric and colorfulness ratings from 8 participants of 30 natural images. In a different approach to measure the perceived color quality of natural images, Hasler and Süsstrunk [15] calculated colorfulness by measuring the color difference against grey. Their metric correlated with participants' ratings of the colorfulness of 84 images at 95% ($r = .95$).

Visual Complexity

While the role of color and colorfulness has been repeatedly named as important for appeal, many more researchers have argued that visual complexity might be the main influence on website appeal [27, 24, 34, 38, 35]. Bauerly and Liu found that complexity in terms of higher numbers of elements on

Table 1. Overview of the implemented image metrics

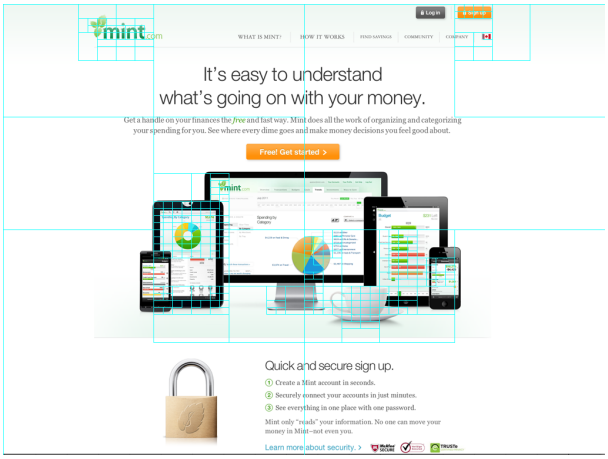
Image metrics	Description
Color	
W3C colors	The percentage of pixels that are close to one of sixteen colors defined by the W3C.
Hue, Saturation, Value	The average pixel value in the HSV color space for hue, saturation, and value, respectively.
Colorfulness [37]	The sum of the average saturation value and its standard deviation where the saturation is computed as chroma divided by lightness in the CIELab color space.
Colorfulness [15]	The weighted sum of the trigonometric length of the standard deviation in ab space and the distance of the center of gravity in ab space to the neutral axis
Space-based decomposition	
Number of leaves	The final number of leaves calculated by the space-based decomposition (modified from [17]), which recursively divides an image into N evenly spaced content regions (leaves), until a region has no visible space divider or until a region is too small.
Number of image areas	Estimates the number of leaves that the algorithm identifies as separate images. Several adjacent images are counted as one image area.
Number of text groups	Refers to the number of horizontal groups of text characters. Each group may represent a word, one-line text, multiple lines of text, or a paragraph.
Text area and non-text area	The number of leaves that have been classified as text or non-text based on a set of heuristics. An example of such heuristic is whether the node has multiple siblings of the same height (an indication that these nodes together are individual characters of the same word).
Quadtree decomposition	
Number of quadtree leaves	Quadtree decomposition using minimum color or intensity entropy as a criterion. Recursively divides an image into subparts until the algorithm converges, and returns a number of leaves (child quadrants) (see [38] for more details).
Symmetry	Evaluates the symmetrical arrangement of the leaves along the horizontal and vertical axes.
Balance	Measures whether the top and bottom, as well as the right and left part of an image have an equal number of leaves, independent of their spatial distribution.
Equilibrium	Evaluates whether the quadtree's leaves mainly center around an image's midpoint.

websites lowers ratings of appeal [1]. Instead, consumers seem to prefer websites that fall within a moderate range of perceived complexity [12]. This is in line with Berlyne's theory [2] that moderate complexity has the biggest arousal potential compared to low or medium levels of complexity. He suggested that the relationship between complexity and appeal represents an inverted U-shape with stimuli that have a low or a high complexity being less preferred than moderately complex ones. Investigating the first impression of a website's complexity and its relation to visual appeal, Tuch et al. [35] recently found an overall linear relationship between the two. In their study, participants showed more negative first impressions of websites with high visual complexity than to those with medium or low complexity.

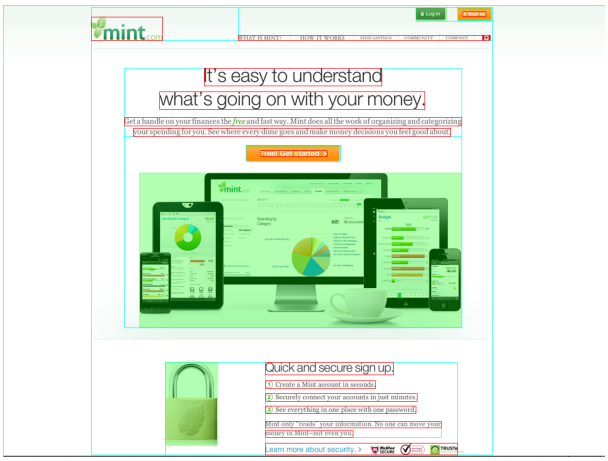
For an explanation of these controversial results, it is necessary to find out which website characteristics contribute to people's perception of complexity. According to Wood [36], people perceive a higher complexity with denser and more dissimilar information presentation. With a similar definition in mind, Rosenholtz et al. [29] predicted whether maps are perceived as cluttered based on a calculation of several features: colors and luminance, different sizes, shapes, and motions. Clutter, in her definition, increases in tandem to the number of unusual objects human attention is drawn to—a concept that is partially related to visual complexity [30]. Visual complexity cannot be measured by the amount of text or the number of images on the user interface alone, but other metrics, such as a largely colorful interface, are thought to additionally contribute to the perception of information overload. For example, a higher color variability (i.e., a large number of colors and a larger distance to each other on the color wheel) has been found to increase the perceived sense of clutter [30].

Previously, research has found that the level of perceived visual complexity can be approximated with a calculation of the percentage of space taken up by text and images based on a calculation of texture [3, 16], by calculating the number of colors [16], or by counting the number of images as defined by a clear boundary [3, 1]. Zheng et al. [38] used the number of leaves resulting from a quadtree decomposition of the image to determine areas of higher information density on webpage screenshots. Their evaluation indicated that the higher the final number of leaves computed by the quadtree decomposition, the more likely websites were to be judged as complicated and unprofessional. A high number of quadrant leaves should therefore also influence the ratings of visual complexity in our study.

Website complexity has also been linked to the structure of a page [3]. Structure describes how the information on an image is spatially distributed, and this can be determined with the help of symmetry, balance, and equilibrium. Symmetry measures how well the left side of an image mirrors the right side, and how well the top mirrors the bottom. In contrast, balance can be achieved in asymmetrical images if both sides are equal in the number of components, and thus, if both sides receive an equal optical weight. Furthermore, equilibrium describes the centering of weight around an image's midpoint.



(a) Quadtree decomposition



(b) Space-based decomposition

Figure 1. The final decomposition result after quadtree and space-based decomposition for an example website.

Symmetry is one of the Gestalt laws' most important predictors for appeal, and it is also an important factor in our aesthetic perception of websites [20, 1]. In Zheng et al.'s evaluation, symmetry and balance significantly correlated with participants' subjective ratings on aesthetics, whereas equilibrium seemed to be largely irrelevant to users' aesthetic judgements [38], probably because websites usually do not have a single focal area in the center.

Computing Procedure

We first selected a set of 450 websites, subdivided into 350 websites in English, 60 foreign websites using a different script, 20 English language websites that were transformed into grayscale images, and 20 websites that had been nominated for the Webby Awards in recent years. All websites were required to have not received wide public exposure and to represent a wide variety of genres.

After obtaining screenshots of each website at a 1024 x 768 pixel resolution, we first computed the percentages of pixels identified as one of the 16 basic W3C colors[1] for each website, as well as the average pixel value for hue, saturation, and value (see Table 1 for a description of all implemented image metrics).

As a next step, histograms for color and intensity were computed as per the procedure in [38]. Color and intensity were measured in the RGB and CIELab color space, respectively. Color entropy and intensity entropy were then calculated from the color/intensity histograms, and used by the quadtree decomposition to decompose our websites into regions of minimum entropy. The quadtree decomposition does this by recursively dividing the image into subparts (the quadrants, or leaves of the tree) along its vertical and horizontal axes. The algorithm computes the entropy of a given criterion (in our case, color or intensity) for each segmented region, and continues to subdivide the region into child quadrants until the entropy drops below a certain threshold. Homogeneously colored regions, for example, have a low color entropy, and thus, the algorithm would not continue to subdivide. Figure 1(a) exemplifies how the quadtree decomposition results in a low number of leaves, because the website has large areas of a uniformly colored background.

Based on the spatial distribution of the resulting leaves of the quadtree decomposition, we further computed symmetry, balance, and equilibrium according to the mathematical definitions provided in [26] and [38].

We additionally added computational metrics based on a space-based decomposition (see Figure 1(b)), which enabled us to identify the outlines of objects on the page, such as text and image areas. Our implementation of the space-based decomposition is similar to the idea of the recursive top-down page segmentation technique X-Y cut as introduced in [17]. In contrast to the quadtree decomposition, space-based decomposition decomposes a page by separating its components along horizontal and vertical *spaces* in the page. The outcome is a tree representing the structure of the website, with the root of the tree being the whole website, the first and second level being main components like the header and the body, and further levels representing subparts of higher levels. The source code for all image metrics can be found on the authors' website at http://iis.seas.harvard.edu/resources/.

EVALUATION OF IMAGE METRICS

While previous implementations of image metrics have rarely been validated in terms of human perception and in the context of websites, this is necessary if we want to be able to make specific inferences about the cause for visual preferences. With Zheng et al.'s work, for example, we are able to say that a specific image metric correlates with people's rating on aesthetics. We are not able to infer that it is the perceived level of complexity, for instance, that led a user to like the page. With the first two experiments, we therefore aimed to collect perceptual human ratings of colorfulness and visual complexity in order to subsequently evaluate which low-level visual features are useful for modeling these perceptions.

Methods

The experiments were implemented as 10-minute online tests on our own experimental platform LabintheWild.org and ad-

[1] `http://www.w3.org/TR/REC-html40/`

vertised in online communities and university newsletters. Both experiments followed the same pattern: Participants were asked to rate screenshots of websites that were shown for 500ms each (following the experiment procedure in [23]). The small exposure time avoids an in-depth engagement with the content of the sites, and instead captures participants' initial reactions towards colorfulness and visual complexity. In the first evaluation phase, we presented each participant with a stratified random sample of 30 websites selected from the larger pool of 450 websites. The 22 english, 4 foreign (using a different script), and 4 grayscale websites were presented in random order. Participants rated every website on a 9-point Likert scale from "not at all complex" to "very complex" or "not at all colorful" to "very colorful", depending on the experiment. After being encouraged to take a short break, we gave a second evaluation phase where participants re-rated the same 30 websites in a different random order so that we could measure consistency in their judgement. Before the two evaluation phases, we gave a short practice phase during which all participants were asked to evaluate a fixed set of five websites, given in randomized order.

We also collected demographic information about each participant, such as gender, age, education level, and current country of residence, in order to control for these factors in the analysis. Participants were excluded from the analysis if they had previously participated in the same study and/or did not have normal or corrected-to-normal vision. Our final data consist of 184 participants (96 female) aged between 15 and 58 years (mean = 21.1) for the colorfulness experiment and 122 participants (60 female) between 16 and 70 years (mean = 32.3) for the complexity experiment.

Data Preparation and Analyses

The standard deviation of the difference between the ratings in phase 1 and 2 across all participants was 0.55 in the colorfulness experiment, and 0.63 in the visual complexity experiment. This indicates a high reliability between participants' ratings in the two test phases, and we therefore used the mean of the two ratings in our analyses.

The slightly higher standard deviation in the complexity experiment might indicate that this concept is more volatile than colorfulness – participants might have based their ratings on slightly different definitions of visual complexity in the two phases. We will later see that this also influences the prediction accuracy of user's perception of visual complexity.

To analyze possible differences between the ratings of different population groups in our sample, we applied two mixed-effects ANOVAs with the demographic variables as fixed effects, StimulusID and ParticipantID as random effects, and the mean rating on colorfulness/complexity as the dependent variable. None of the demographic variables (country of current residence, gender, age, and education level) had a significant main effect on mean colorfulness or mean complexity, suggesting that the perception of colorfulness and visual complexity is more or less universal. People as a whole seem to make very similar judgements on these website characteristics, independent of their demographic background.

Table 2. Regression model for the perception of colorfulness, $R^2 = .78, p < .001$

	b	SE b	β	$p <$
Constant	-.68	.73		.05
Gray	2.45	.56	.24	.001
White	2.74	.71	.44	.001
Maroon	1.60	.71	.20	.05
Green	1.85	.57	.23	.01
Lime	-3.17	1.42	-.06	.05
Blue	-3.91	1.06	-.11	.001
Teal	1.01	.56	.09	.05
Saturation	.01	.002	.13	.01
Colorfulness [15]	.03	.003	.58	.001
Number of image areas	.06	.008	.19	.001
Number of quadtree leaves	3.74E-4	.000	.153	.001
Text area	-1.48E-6	.000	-.06	.05
Non text area	1.86E-6	.000	.192	.001

The 20 grayscale website screenshots received an average rating of 1.22 (sd = 0.07), demonstrating that they were correctly identified as colorless. They were excluded from the subsequent analysis since the image metrics dependent on color information cannot be reliably computed for these screenshots. We additionally excluded websites that had received three or fewer ratings (due to the random assignment to participants), or where the standard error of their mean complexity was ≥ 0.75. Subsequent analyses therefore report on 421 webpages for the analysis of colorfulness, and 382 for the analysis of complexity (out of 450 websites).

To determine the most predictive image metrics for the ratings on both concepts, we used multiple linear regression with backward elimination. In this method, all predictors are initially added to the model, and iteratively removed if they do not make a statistically significant contribution to how well the model predicts the outcome variable (the ratings, in our case). At each step, the remaining predictors are reassessed in terms of their contribution to the newly calculated model.

Prediction Model for Perceived Colorfulness

Based on the related literature on methods for calculating perceived colorfulness, we had seven image metrics plus the 16 different HTML color percentages in our initial regression model.

We found that both colorfulness metrics by Yendrikhovskij et al. [37] and Hasler and Suesstrunk [15] significantly correlate with the mean ratings on perceived colorfulness ($r = .53$ and $r = .71$, $p < .001$). This shows that although they were initially meant to approximate humans' colorfulness judgments in natural images (where they achieved much higher prediction accuracies, see [37] and [15]), they also serve as solid predictors for website colorfulness. However, because these two metrics are also highly correlated with each other, the colorfulness metric by Yendrikhovskij et al. was ultimately removed from the model during the stepwise procedure due to the presence of Hasler's metric.

It is interesting to note that the previous colorfulness metrics do not consider hue, yet our regression model considers sev-

eral color percentages as valuable for predicting people's perception of colorfulness. Moreover, the final regression model (see Table 2) also acknowledges the influences of the number of image and text areas on humans' perceived colorfulness, despite the fact that the colorfulness of these areas alone has already been taken into account by the other metrics.

The final regression model is highly predictive of participants' mean ratings of colorfulness ($r = .88$, $p < .001$, $R^2 = .78$, adj. $R^2 = .77$). The predicted ratings using this model are shown in Figure 2(a).

There are isolated outliers where participants' perception of colorfulness did not match the prediction of our model. The website in Figure 3(a), for example, only received a mean colorfulness rating of 3.11. Our model, which interprets large areas of highly saturated colors as more colorful, overestimated this rating by 2.8 points. Figure 3(b) additionally shows an example where our model underestimated the perceived colorfulness of 5.46 by close to two points. Although the website screenshot is mainly white, participants were apparently substantially affected by the contrasting, brightly colored buttons in the middle of the screen. The same screenshot received a visual complexity rating of 2 on average, which suggests little contribution of the colorful buttons to the perception of complexity.

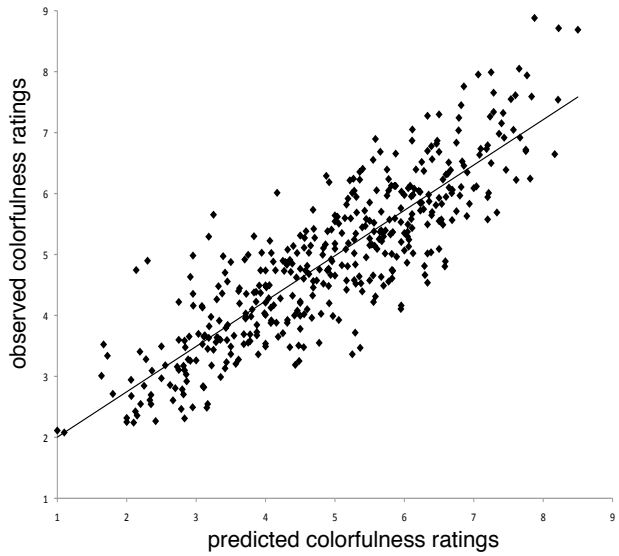
Prediction Model for Perceived Visual Complexity

While Zheng et al's evaluation [38] suggested that a higher number of quadtree leaves is related to perceiving a website as more complicated, our results show that this concept might only partially contribute to perceived visual complexity. In fact, the correlation of the mean rating on visual complexity with the number of quadtree leaves is weak ($r = .28$, CI = .18–.37, N = 367, level = .95). With a correlation coefficient of $r = .5$ (CI = .42–.57, N = 367, level = .95), the number of leaves calculated by the space-based decomposition seems to be more representative of perceived visual complexity. As we saw in Figure 1(b), this approach does indeed intuitively make more sense as a way to represent the number of objects on a page, which, according to [36] and [29], plays an important role in the related concept of clutter [30].

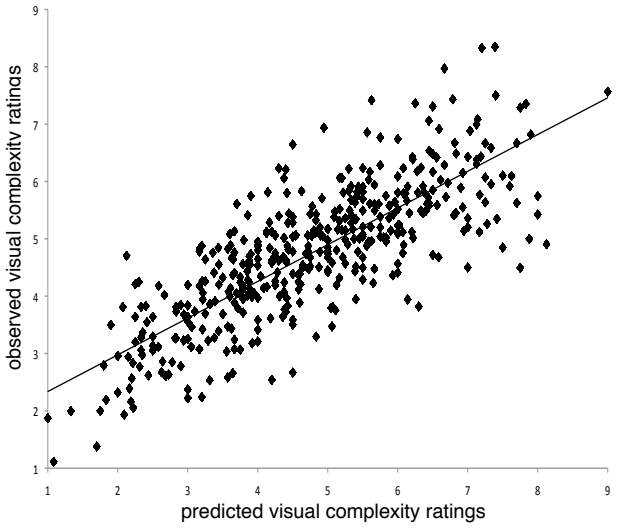
Table 3. Final regression model for the perception of visual complexity, $R^2 = .65, p < .001$

	b	SE b	β	$p <$
Constant	.637	.179		.001
Text area	.000	.000	.407	.001
Non text area	.000	.000	.515	.001
Number of leaves	.005	.002	.123	.05
Number of text groups	.052	.008	.344	.001
Number of image areas	.056	.012	.193	.001
Colorfulness [37]	.011	.005	.079	.05
Hue	.005	.002	.073	.05

Due to Zheng et al.'s work [38], we had additionally assumed that the spatial metrics computed with the help of the quadtree decomposition (balance, symmetry, and equilibrium) would contribute to the perception of orderliness and visual complexity. These metrics showed weak but significant correlations with the mean complexity ratings, but were ultimately pruned from the model.



(a) Colorfulness with a model fit of $R^2 = .78$, $p < .001$

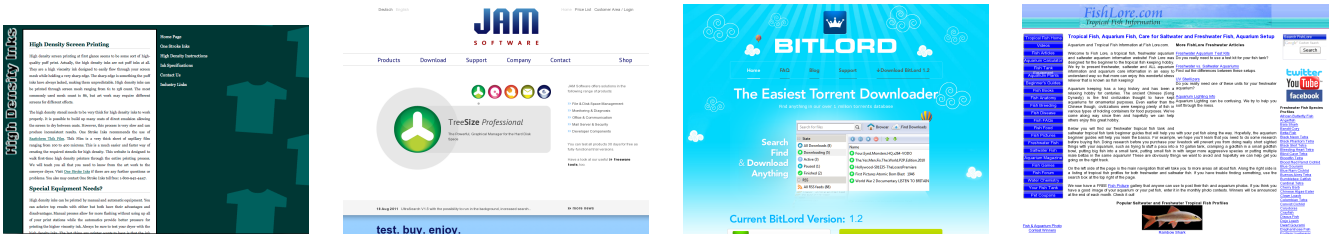


(b) Visual complexity with a model fit of $R^2 = .65$, $p < .001$

Figure 2. Mean colorfulness/visual complexity ratings vs. predicted from the respective multiple regression equations.

While the colorfulness metric developed by Yendrikhovskij et al. [37] is included in the model, saturation and value are were removed from the final model. As mentioned in the previous section, Yendrikhovskij et al's colorfulness metric includes saturation, and so its exclusion does not manifest its irrelevance to people's perception of visual complexity.

The final model as shown in Table 3 is highly correlated with participants' mean ratings on visual complexity ($r = .80$, $p < .001$, $R^2 = .65$, adj. $R^2 = .64$). However, Figure 2(b) also shows that visual complexity is a slightly more volatile concept than colorfulness. In fact, the slight outliers reveal that our computational model seems to overestimate the in-



(a) Colorfulness rating 3.11 vs. a predicted rating of 5.93.

(b) Colorfulness rating 5.46 vs. a predicted rating of 3.75.

(c) Complexity rating 2.13 vs. a predicted rating of 4.7.

(d) Complexity rating 5.63 vs. a predicted rating of 7.42.

Figure 3. Examples of over- and underestimation of perceived colorfulness and complexity ratings by our model.

fluence of bright colors (see Figure 3(c)). In addition, Figure 3(d) illustrates an overestimation of the influence of text; while our model predicted a rating of 7.42 for this text-heavy website, participants actually perceived it as significantly less complex (5.63).

The high prediction accuracy of both models allowed their use in our next step, the prediction of users' first impression of appeal. The following section describes the experiment and regression model for the prediction of visual appeal.

EXPERIMENT ON VISUAL APPEAL

In the previous two experiments we established that our computational models are effective in predicting a website's visual complexity and colorfulness as perceived by users. The goal of this third experiment was to evaluate whether our models of perceived colorfulness and complexity can indeed serve as predictors for appeal.

Method

We used the same online set-up as in the previous two experiments with the exception that participants were asked to rate the website screenshots on their perceived visual appeal.

Because visual preferences have previously been found to differ by gender [7, 33], education level [4], or between countries [8, 28], we tried to attract a diverse audience in order to collect data from participants with possibly heterogeneous visual preferences. As an incentive, participants received feedback on their own visual taste at the end of the study, and they were able to compare their preference for a certain level of colorfulness or visual complexity to those of people from other countries.

The experiment was conducted on our experimental platform LabintheWild.org, and completed by 242 volunteers (103 female) who reported that they had normal or corrected-to-normal vision. They represented a large variety of backgrounds, ranging between 16 and 70 years in age (mean = 32.3 years), and living in 34 different countries. The majority of participants (28.5%) came from the US, and an additional 31.8% participants currently lived in the US but had lived in other countries before. We additionally collected information about native language, education level, Internet usage on a normal day, and profession.

Data Preparation and Analyses

We first analyzed whether participants' ratings were consistent across phase 1 and 2, and excluded the 218 out of 3630 rating pairs with deviations of more than 2 points on the 9-point Likert scale. This ensured that we only included those ratings that are representative of a participant's preferences. The following results therefore report on 3412 rating means. The standard deviation of the differences of ratings in the final pairs was 0.5.

Results

To analyze the influence of predictor variables on the outcome of participants' ratings on visual appeal, we used a linear mixed-effects model, where StimulusID and ParticipantID were modeled as random effects, and complexity, colorfulness, and an interaction between all demographic variables and the complexity and colorfulness (using the previously constructed models) were modeled as fixed effects. Quadratic terms for complexity and colorfulness (and demographic interactions with these terms) were also included to allow for the postulated "U-shape" relationship discussed earlier.

The model accounts for 48% of the variation in aesthetic preferences (adj. $R^2 = .48$, see also Figure 4). This means that in combination with demographic variables, users' first impression of a site is highly explained by the sites' perceived colorfulness and visual complexity alone. The histogram of residuals (i.e., the individual-level errors) indicated normality. Coefficient estimates, standard errors, and significance levels for all predictors in the analysis can be accessed on the author's website at http://iis.seas.harvard.edu/resources/.

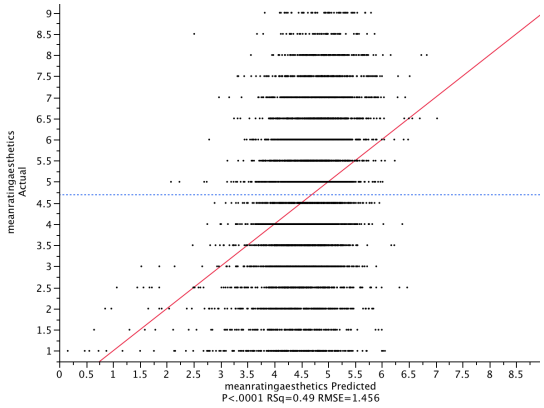


Figure 4. A visualization of the model of visual appeal using individual ratings as the outcome variable. Red line shows the fit, and the blue line indicates the mean rating across participants.

The Impact of Colorfulness and Complexity

Our results show that ratings of appeal are significantly negatively affected by an increase in visual complexity. As shown in Figure 5, we observed a strong decrease in ratings of appeal for websites with a high level of complexity, as well as a slight decline in the ratings for websites with lower levels of complexity. This is also in line with the results shown in [35].

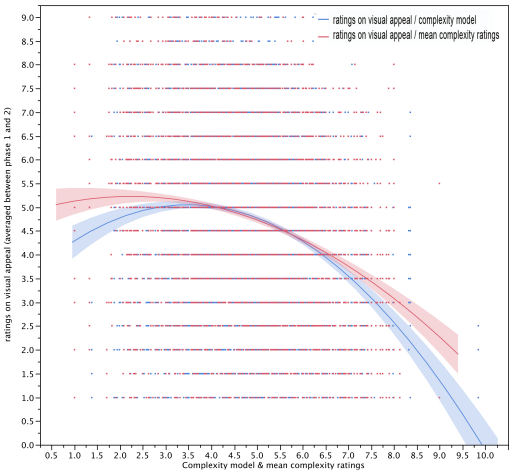


Figure 5. The ratings on visual appeal in relation with our complexity model (blue line) and the mean complexity ratings by our participants (red line).

The Influence of Demographic Background

We found a significant main effect of age on the outcome variable ($F_{(4)} = 6.68$, $p < .001$), but not for education level and gender ($p > .20$ and $p > .15$, respectively). Because many of our participants have lived in multiple countries, we also evaluated the effect of native language to test whether a participant's origin has a predictive effect on aesthetic rating. Native language had a significant main effect on the ratings of appeal ($F_{(26)} = 2.4$, $p < .001$), but did not substantially improve the model fit. Hence, we removed native language from the final model. Because of the low number of participants per country/native language, we believe that these results deserve a more in-depth analysis with a larger sample in the future.

Looking at the interaction terms of demographic variables with colorfulness and complexity, we see that the contribution of colorfulness and complexity on people's aesthetic impression is not universal. Age did not significantly interact with colorfulness, but with complexity ($F_{(4)} = 6.53$, $p < .001$). In particular, participants older than 45 years liked websites with a low visual complexity level more than other age groups.

We further observed an interaction effect of education level with colorfulness ($F_{(4)} = 2.61$, $p < .05$). Participants with a PhD were most negatively affected by a high colorfulness, although those participants with a high school education preferred websites with a similarly low colorfulness. Gender did not show significant interaction effects with colorfulness and complexity.

What is interesting about the interaction between some of the demographic variables and colorfulness /complexity is that it is often assumed that we can find design guidelines for a universal audience. For example, it is often assumed that the Webby Award websites are universally liked, and it is common to use these websites as best practice examples [16, 38, 19]. Yet the 20 Webby Award websites in our study received average ratings between 4.21 and 6.57 (across all participants on a 9-point scale, and as compared to the overall mean of 4.73 for all the other sites). A relatively high average standard deviation of 1.69 across all participants and Webby Award sites indicates a high dispersion of participants' preferences for these sites. When excluding those participants from the analysis who had lived in places other than the US, and/or had parents of a different nationality, we see only slightly better ratings: Americans rated the Webby Award websites between 5.15 and 6.39 on average, and the average standard deviation across all of the US participants was 1.65. The results emphasize that aesthetic preferences at first sight differ even for supposedly well-designed websites.

DISCUSSION

Our findings support the strong role of visual complexity in users' first impressions of appeal, in line with much previous work [24, 38, 35]. In particular, our results correspond to those presented in a recent paper by Tuch et al. [35], who found that the relationship between visual complexity and appeal is not necessarily represented by a strong U-shape. In fact, our analyses show that a high visual complexity results in the largest decrease in appeal, but websites with low levels of complexity are similarly liked to those with a medium complexity. Extending previous work, our models now provide a computational way to assess the influence of visual complexity on visual appeal requiring only a website screenshot as input.

An interesting finding of our work is that the perceived colorfulness only plays a minor role in people's first impression of appeal. To interpret this result, we need to take a step back and look at our model of perceived visual complexity, which already partly accounts for the influence of colors. However, our two perceptual models only weakly correlate with each other. This suggests that, while slightly interacting, the models do indeed measure different website characteristics. Our result therefore suggests that perceived visual complexity is the more salient website feature for visual appeal at first sight.

We also found that there is a fair amount of wisdom in the popular saying "Beauty is in the eye of the beholder." Preferences vary, with even Webby Award websites being disliked by many participants. This suggests the need for personal models of appeal. In our analysis, treating participants as a random variable and including interaction effects between the complexity and colorfulness models and the demographic variables age group, education level, and gender resulted in the best model fit. In particular, we found that education level significantly interacted with colorfulness, and age showed strong interaction effects with complexity. However, unlike in the findings presented in [7], gender did not show any significant interaction effects with colorfulness or complexity.

The final model explains 48% of the variance in participants' ratings on appeal formed after 500ms of viewing time. A

website's visual complexity and colorfulness seem to explain much of a user's initial reaction towards a website's appeal.

LIMITATIONS AND FUTURE DIRECTIONS

When we began our studies of websites aesthetics, our goal was to predict people's first impression of appeal. Hence, this work was not intended to predict judgements that have been found to be influenced by appeal, such as the perceived trustworthiness and usability. Similarly, our findings may not generalize to predict user's "long-term" appeal as conscious and careful analysis of a website over a longer time might change a user's opinion. We believe that the question on how we can quantify such aesthetic change over time is worth pursuing in future work.

A limitation of our study is that the 450 evaluated websites may not represent a random sample drawn from the Internet. Although we carefully constructed this dataset to be as representative of the real world as possible, we cannot claim that it captures the same aesthetic diversity as found on the Web. Further studies with larger datasets will be necessary to validate the generalizability of our models to the real world.

Along these lines, our experiments involved samples of more than 500 participants, yet this sample does not necessarily represent the wide spectrum of typical Internet users. In fact, the demographics information that we collected revealed that our sample is more educated and more multinational than typical Internet users. More homogeneous sub-samples and an overall larger number of participants are necessary to turn the small demographic effects that we found into definite statements.

An obvious direction for future work is also the inclusion of more aesthetic image metrics in order to evaluate what other characteristics account for the variation in users' ratings on appeal. Moreover, we are excited to utilize our models for an automatic adaptation of websites to suit users' personal preferences. Our next steps therefore include the implementation of tools that automatically rearrange websites to fit user-specified levels of colorfulness and visual complexity.

CONCLUSION

Although it is generally uncontested that also for websites "the first impression counts," research has mostly concentrated on providing qualitative design guidelines to improve users' perception of appeal. In this paper, we presented quantitative models for the prediction of appeal, thus enabling an automatic judgment of designs.

We first introduced a new model of perceived colorfulness and visual complexity, developed based on the ratings of 306 participants. We demonstrated that users' initial perception of colorfulness and complexity can be quantified with the help of low-level image features of static website screenshots. This provides new methods for designers to judge and compare website designs, but it also allowed us to take the next step towards our goal of predicting users' first impressions of visual appeal.

To this end, we asked 242 participants to rate an overall set of 450 websites on visual appeal. Building on our previously established models, we demonstrated that—in combination with demographic variables—colorfulness and visual complexity explain 48% of the variance in users' first impressions. Our results show that visual complexity accounts for a significantly larger amount of the variance in the observed ratings on visual appeal than colorfulness. Moreover, our findings suggest that the importance of these two website characteristics is not universal, but dependent on users' demographic backgrounds. Our results pave the way for larger endeavors to improve the user experience on the web, because the first impression counts.

ACKNOWLEDGMENTS

This work was funded in part by a Harvard Mind/Brain/Behavior Faculty Award. Katharina Reinecke was supported by the Swiss National Science Foundation under fellowship number PBZHP2-135922.

REFERENCES

1. Bauerly, M., and Liu, Y. Effects of Symmetry and Number of Compositional Elements on Interface and Design Aesthetics. *Int. Journal of Human-Computer Interaction 3* (2008), 275–287.
2. Berlyne, D. *Studies in the New Experimental Aesthetics*. Washington, DC: Hemisphere Pub. Corp., 1974.
3. Bucy, E. P., Lang, A., Potter, R. F., and Grabe, M. E. Formal Features of Cyberspace: Relationships Between Web Page Complexity and Site Traffic. *Journal of the American Society for Information Science 50*, 13 (1999), 1246–1256.
4. Chen, J. Y., Whitfield, T. W. A., Robertson, K., and Chen, Y. The Effect of Cultural and Educational Background in the Aesthetic Responses of Website Users, 2010. National Institute for Design Research, Swinburne University of Technology, AU.
5. Coursaris, C., Swierenga, S., and Warall, E. An Empirical Investigation of Color Temperature and Gender Effects on Web Aesthetics. *Journal of Usability Studies 3*, 3 (2008), 103–117.
6. Cyr, D. Modeling Website Design across Cultures: Relationships to Trust, Satisfaction and E-loyalty. *Journal of Management Information Systems 24*, 4 (2008), 47–72.
7. Cyr, D., and Bonanni, C. Gender and Website Design in e-Business. *International Journal of Electronic Business 3*, 6 (2005), 565–582.
8. Cyr, D., Head, M., and Larios, H. Colour Appeal in Website Design Within and Across Cultures: A Multi-method Evaluation. *Int. Journal of Human-Computer Studies 68*, 1-2 (2010), 1–21.
9. Datta, R., Joshi, D., Li, J., and Wang, J. Studying Aesthetics in Photographic Images Using a Computational Approach. In *Proc. Computer Vision–ECCV* (2006), 288–301.
10. Desnoyer, M., and Wettergreen, D. Aesthetic Image Classification for Autonomous Agents. In *Proc. Pattern Recognition* (2010), 3452–3455.

11. Everard, A., and Galletta, D. How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store. *Journal of Management Information Systems 22*, 3 (2006), 5595.

12. Geissler, G., Zinkhan, G., and Watson, R. The Influence of Home Page Complexity on Consumer Attention, Attitudes, and Purchase Intent. *Journal of Advertising 35*, 2 (2006), 69–80.

13. 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.

14. Hartmann, J., Sutcliffe, A., and Angeli, A. D. Towards a Theory of User Judgment of Aesthetics and User Interface Quality. *ACM Transactions on Computer-Human Interaction 15*, 4 (2008).

15. Hasler, D., and Suesstrunk, S. Measuring Colourfulness in Natural Images. In *Proc. SPIE/IS&T Human Vision and Electronic Imaging*, vol. 5007 (2003), 87–95.

16. Ivory, M., Sinha, R., and Hearst, M. Empirically Validated Web Page Design Metrics. In *Proc. CHI* (2001), 53–60.

17. Jaekyu Ha, Haralick, R., and Phillips, I. Recursive XY Cut using Bounding Boxes of Connected Components. In *Proc. Document Analysis and Recognition* (1995), 952–955.

18. Kim, J., and Moon, J. Y. Designing Towards Emotional Usability in Customer Interfaces - Trustworthiness of Cyber-banking System Interfaces. *Interacting With Computers 10* (1998), 1–29.

19. Kumar, R., Talton, J. O., Ahmad, S., and Klemmer, S. R. Bricolage: Example-based Retargeting for Web Design. *Proc. CHI'11* (2011), 2197–2206.

20. Lavie, T., and Tractinsky, N. Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. *International Journal of Human-Computer Studies 60*, 3 (2004), 269–298.

21. Lindgaard, G. Aesthetics, Visual Appeal, Usability and User Satisfaction: What Do the User's Eyes Tell the User's Brain? *Australian Journal of Emerging Technologies and Society 5*, 1 (2007), 1–14.

22. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. An Exploration of Relations Between Visual Appeal, Trustworthiness and Perceived Usability of Homepages. *ACM Transactions on Computer-Human Interaction 18*, 1 (2011).

23. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression! *Behaviour & Information Technology 25*, 2 (2006), 115–126.

24. Michailidou, E., Harper, S., and Bechhofer, S. Visual Complexity and Aesthetic Perception of Web Pages. *Proc. Design of Communication* (2008), 215–224.

25. Moshagen, M., Musch, J., and Göritz, A. S. A Blessing, not a Curse: Experimental Evidence for Beneficial Effects of Visual Aesthetics on Performance. *Ergonomics 52*, 10 (2009), 1311–1320.

26. Ngo, D., Teo, L., and Byrne, J. Modelling Interface Aesthetics. *Information Sciences* (2003), 25–46.

27. Pandir, M., and Knight, J. Homepage Aesthetics: The Search for Preference Factors and the Challenges of Subjectivity. *Interacting With Computers 18*, 6 (2006).

28. Reinecke, K., and Bernstein, A. Improving Performance, Perceived Usability, and Aesthetics with Culturally Adaptive User Interfaces. *ACM Transactions on Computer-Human Interaction (ToCHI) 18*, 2 (2011).

29. Rosenholtz, R., Li, Y., Mansfield, J., and Jin, Z. Feature Congestion: A Measure of Display Clutter. In *Proc. CHI* (2005).

30. Rosenholtz, R., Li, Y., and Nakano, L. Measuring Visual Clutter. *Journal of Vision 7*, 2 (2007).

31. Sonderegger, A., and Sauer, J. The Influence of Design Aesthetics in Usability Testing: Effects on User Performance and Perceived Usability. *Applied Ergonomics*, 41 (2010), 403–410.

32. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages. *Int. Journal of Human-Computer Studies 64* (2006).

33. Tuch, A. N., Bargas-Avila, J. A., and Opwis, K. Symmetry and Aesthetics in Website Design: It's a Man's Business. *Computers in Human Behavior 26*, 6 (2010), 18311837.

34. Tuch, A. N., Bargas-Avila, J. A., Opwis, K., and Wilhelm, F. H. Visual Complexity of Websites: Effects on Users' Experience, Physiology, Performance, and Memory. *Int. Journal of Human-Computer Studies 67* (2009), 703–715.

35. Tuch, A. N., Presslaber, E., Stoecklin, M., Opwis, K., and Bargas-Avila, J. The Role of Visual Complexity and Prototypicality Regarding First Impression of Websites: Working Towards Understanding Aesthetic Judgments. *International Journal of Human-Computer Studies 70(11)* (2012), 794–811.

36. Wood, R. E. Task Complexity: Definition of a Construct. *Organizational Behavior and Human Decision Processes 31*, 1 (1986), 60–82.

37. Yendrikhovskij, S. N., Blommaert, F. J. J., and De Ridder, H. Optimizing Color Reproduction of Natural Images. In *Proc. Color Imaging Conference: Color Science, Systems, and Applications* (1998), 140–145.

38. Zheng, X., Chakraborty, I., Lin, J., and Rauschenberger, R. Correlating Low-level Image Statistics with Users' Rapid Aesthetic and Affective Judgments of Web Pages. In *Proc. CHI* (2009).