A thesis submitted to the University of Trento for the degree of Doctor of Philosophy

Computational Aesthetics in HCI: Towards a Predictive Model of Graphical User Interface Aesthetics.

Aliaksei Miniukovich

Advisor: Prof. Dr. Antonella De Angeli

International Doctoral School in Information and Communication Technologies
University of Trento

Statement of Contribution
This thesis reports research principally done by the author, as a part of his doctoral research.
This thesis consists of the following publications:

Chapter 2.
Miniukovich, A., & De Angeli, A. (2014, May). Quantification of interface visual complexity. In *Proceedings of the 2014 international working conference on advanced visual interfaces* (pp. 153-160). ACM.

Miniukovich, A., & De Angeli, A. (2014, October). Visual impressions of mobile app interfaces. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (pp. 31-40). ACM.

Miniukovich, A., & De Angeli, A. (2015, April). Computation of interface aesthetics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1163-1172). ACM. **The paper has been recognized with an honorable-mention award.**

Chapter 3.
Miniukovich, A., & De Angeli, A. (2016). Computational Aesthetics: From Webpages to Websites. *Under review at ACM TOCHI.*

Miniukovich, A., & De Angeli, A. (2015, July). Visual diversity and user interface quality. In *Proceedings of the 2015 British HCI Conference* (pp. 101-109). ACM.

Miniukovich, A., & De Angeli, A. (2016). Pick Me! Getting Noticed on Google Play. *To appear in Proceedings of the 34rd Annual ACM Conference on Human Factors in Computing Systems.* ACM. **The paper has been recognized with an honorable-mention award.**

Table of Contents

1 INTRODUCTION

1.1 Introduction

This thesis describes the development and validation of a predictive model of graphical user interface (GUI) aesthetics. The development was informed by the processing-fluency theory of aesthetic pleasure and involved outlining several visual dimensions of GUI designs, which could affect aesthetics impression. Each of the dimensions was grounded in theory and represents a unique visual aspect of GUI design. The resulting model automatically evaluates the design dimensions and combines them in an estimate of the average impression that GUI appearance would make on the user population. The model was validated in a number of user studies proving high validity and reliability. The model outputs an aesthetics score of user impression and could inform the creation of more beautiful GUIs by highlighting which of the design dimensions could be improved. The thesis describes the studies that validated the model on several types of GUIs and demonstrated a potential application of the model in future research and practice.

1.1.1 Motivation

This thesis presents a predictive model of visual aesthetics for graphical user interfaces. Aesthetics corresponds to the perception-based impression of GUI visual appearance, which influences the user appreciation and attitude towards the GUI. The model analyzes GUIs and forecasts the scores that the user would give to the aesthetics of the GUIs. The ability to forecast differentiates the model from past work, which proposed descriptive models summarizing and structuring aesthetics-related phenomena (Tractinsky, 2013; Hassenzahl, 2005; Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010). Ten user studies have tested the validity and reliability of the model on various GUIs, such as websites, mobile apps and app icons.

HCI work has recognized visual aesthetics as a distinct and impactful component of GUI quality (Kurosu & Kashimura, 1995; Tractinsky, 1997). A relevant corpus of research has discussed the properties of aesthetics (Tractinsky et al., 2006; Lindgaard et al., 2006; 2011) and its interdependence with other UX dimensions (Tractinsky et al., 2000; De Angeli et al., 2006; Sonderegger & Sauer, 2010). Some scholars have developed multi-item measurement scales for aesthetics (Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010), positioned it relative to the other quality components in UX models (Hassenzahl, 2004; Van Schaik & Ling, 2008; Hartmann et al., 2008), speculated about its antecedents (Hall & Hanna, 2004; Bauerly & Liu, 2006; Tuch et al., 2009), outcomes (Cyr et al., 2010; Braddy et al., 2008; Thielsch et al., 2013) and its decreasing impact on UX as usage time passed (Karapanos et al., 2009; Sonderegger et al., 2012). However, the outcomes of these investigations still struggled to directly inform the creation of aesthetical GUIs, which appeared a substantial challenge requiring a solution (Hassenzahl, 2012).

As a possible solution, the aesthetics design guidelines (Sutcliffe, 2009; 2002; Chang et al., 2002) tried to extrapolate design principles from the psychological knowledge on human perception. However, such knowledge was derived from experimenting with the sets of polygons, which is typical in psychology research, and not from experimenting with realistic GUIs. These guidelines gave recommendations either too general to be useful (e.g., in Sutcliffe, 2009, "*we naturally see the complete object such as a circle*" tells little about beauty,), or too specific to be non-controversial and to apply in the multitude of real-world situations. Thus, "*low saturation pastel colours should be used for backgrounds*" (Sutcliffe, 2009) contradicts Lindgaard et al.'s (2011, p. 20) observation of high-aesthetics pages tending to have bright and saturated backgrounds. The low-saturation pastels may fit some situations, but not all of them.

This thesis claims that computational aesthetics can tackle the issue of designing appealing GUIs more effectively than the design guidelines can. The increased effectiveness is due to several reasons. First, computational aesthetics can give non-generic recommendations: it uses precisely-defined design dimensions – distinct visual aspects of design, such as the number of colors or number of layout columns – and links the dimensions to aesthetics unambiguously, decreasing the argument amongst evaluators if, for example, a design is colorful or non-colorful and if such colorfulness decreases or increases aesthetics. Second, computational aesthetics can give recommendations in a wide range of situations: it extracts the design dimensions from analyzing *large* stimuli samples, far larger than the samples used in similar manual explorations of designs (e.g., Kim et al., 2003; Park et

al., 2004 or study 3 in Lindgaard et al., 2006). Including more samples in an analysis extends the validity and reliability of results. Third, computational aesthetics can rely on a continuous score to describe a design, whereas the guidelines often assume a dichotomy, e.g., a design is categorized as either symmetrical or asymmetrical, which oversimplifies real-world GUIs. Finally, computational aesthetics can give recommendations that require less expertise to interpret than the guidelines: the aesthetics-estimation algorithms can visualize design dimensions, which substantially facilitates understanding design and making design improvements (Rosenholtz et al., 2011).

A small corpus of recent research had targeted developing the methods of aesthetics computation in HCI, and the methods had shortcomings. Some of them did not rely on a theory to motivate the identification of design dimensions or the link between the dimensions and aesthetics (e.g., Datta et al., 2006). Instead, these methods relied on examining countless GUI features, which might have happened to correlate with aesthetics scores, but lacked a theoretical backing to substantiate the correlation. As an illustrative example, webpage appreciation happened to depend on webpage meta keywords, and number of scripts and applets embedded in a webpage (Ivory et al., 2001; Ivory & Hearst, 2002) or coarseness, contrast and directionality of webpage textures (Wu et al., 2011). Because the methods used a *large* number of features and statistical methods suitable for the many-feature situations (e.g., the classification and regression tree analysis in Ivory & Hearst, 2002, or support vector machines in Wu et al., 2011), they could not result in *simple* explanations on how aesthetics could be improved. Any such explanation would require an expert to interpret the outcome of algorithms. The past attempt to build easy-to-understand good-webpage profiles (Ivory & Hearst, 2002) gave confusing results (e.g., having an italicized word on a webpage happened to attribute the webpage to the poor-design category).

Other computational methods (Ngo et al., 2003; Ngo, 2001; Ngo et al., 2000) did rely on the theories of design aesthetics from visual arts (e.g., Arnheim, 1954 as cited in Ngo et al., 2000) and leveraged relatively few features. However, the attempts to validate these methods had produced limited and contradictory evidence. Only few of Ngo's computational measures (Ngo, 2001; Ngo et al., 2000; 2003) produced scores correlating with aesthetics scores and these few successful measures differed across different studies. For example, Purchase et al. (Purchase et al., 2011) tested each of 14 of Ngo's aesthetics measures on just 15 datapoints and found only few of the measures to correlate with webpage aesthetics (for the color versions of webpages, these measures included proportion, cohesion and balance). Zheng et al. (2009) tested only webpage balance, symmetry, and equilibrium. Equilibrium did not correlate with webpage aesthetics. Altaboli et al. (2011) also tested three measures – webpage balance, unity and sequence – and found unity and sequence to correlate with aesthetics. Balance did not correlate with aesthetics, which, again, contradicted the other two studies and the theory.

This thesis addresses the shortcomings of the past methods for GUI aesthetics computation. It leverages the processing fluency theory (Reber et al., 2004) to drive the selection of design dimensions and user studies to validate the relationship between the dimensions and GUI aesthetics. The outlined dimensions constitute a major theoretical contribution of the thesis. The algorithms to estimate the dimensions constitute a major practical contribution of the thesis. The work has resulted in a predictive model of GUI aesthetics, which has been validated on several types of GUIs.

1.1.2 Theory

Two main epistemological positions underlie aesthetics research in HCI. One position is grounded in psychology and considers aesthetics a visceral reaction (e.g., Tractinsky, 2013; Lindgaard et al., 2006; De Angeli et al., 2006); the other position is grounded in fine arts and considers aesthetics an experience (e.g., Wright et al., 2008; Blythe et al., 2010). The psychological position focuses on user's impression of a stimulus; the artistic position focuses on user's interpretations of the stimulus. The psychological position measures the semi-automatic, quickly formed and culture-independent appreciation of stimuli, and considers the high consistency in users' reactions in the same context as a sign of validity. The artistic position collects the conscious, reason-based and culture-dependent judgments of stimuli, and considers the diversity in users' interpretations in different contexts as a sign of validity. The psychological position decomposes aesthetics into constituents and studies them separately in controlled experiments. The artistic position studies aesthetics as a holistic concept inseparable from the context it has occurred into. Both positions might have useful applications,

however only the psychological position allows for building predictive models: the artistic position relies on cases studies and post-hoc analyses of experiences, which can only result in descriptive models. This thesis aimed at building a predictive model of aesthetics, and thus, has stayed within the psychological position.

Within the psychological position, different theories of aesthetics exist, which could provide the rationale for the selection of aesthetics design dimensions. Berlyne's collative theory (Berlyne, 1971) tried to explain aesthetics preferences as an outcome of "pleasingness", which in turn depended via an inverted U-shaped link on the collative variables, such as uncertainty, complexity or surprisingness. All collative variables were presumed to measure the physiological arousal of organisms due to new incoming information mismatching existing knowledge or expectations. Berlyne's models could not explain a large amount of empirical evidence and should no longer be considered valid (Martindale et al., 1990). A prototype-preference theory (Whitfield, 1983) described the aesthetics preference as an outcome of prototypicality: higher resemblance to a prototype would lead to higher aesthetical appreciation. This theory also failed to explain some empirical evidence, e.g., the peak-shift effect (a preference for the stimuli that resemble a prototype, but also include a dose of novelty, (Ramachandran & Hirstein, 1999; Hekkert et al., 2003)). The processing fluency theory (Reber et al., 2004; Reber, 2012; Winkielman et al., 2002) explains aesthetics preferences as a function of how fluently a mind can process a stimulus: higher fluency leads to a lower mental effort and higher aesthetic appreciation. The processing of stimuli in the mind happens subconsciously and very quickly, which explains the recent HCI findings suggesting that stable aesthetics preferences form very quickly (Lindgaard et al., 2006; 2011; Tractinsky et al., 2006). The processing-fluency theory does not account for the preferences that conscious reasoning produces, and thus, could only explain the subconscious-based part of aesthetic pleasure (Armstrong & Detweiler-Bedel, 2008; Silvia, 2012). Nonetheless, the subconscious part appeared to be universal (Reber, 2012) and was easier to associate with design dimensions than the reasoning-based part that heavily depended on user's culture, education, personal experience and other variables that are difficult to control for. The thesis has relied on the theory of processing fluency to guide the development of design dimensions that could predict design aesthetics.

The applications of processing-fluency theory to aesthetics (Reber, 2012) built on the description of aesthetics attributes by the philosopher George Santayana (1955): aesthetics is a *value-positive*, *intrinsic*, and *objectified* concept (cf., Moshagen & Thielsch, 2010). The attribute of *value-positive* implies that aesthetics results in pleasure, leaving unpleasantness (e.g., due to the possible provocative nature of artworks) outside of this thesis focus. The attribute of *intrinsic* implies that aesthetics results in preferences almost immediately, without the interference of elaborate reasoning about the standards of beauty or usefulness. Such elaborate standards cannot be inborn; they are learnt. The non-reliance on elaborate reasoning and learnt standards implies that everyday objects – including GUIs – can be intrinsically beautiful. The attribute of objectified implies aesthetics perception is directed towards an object, rather than towards body sensations: enjoying the warmth of hot coffee does not make the coffee beautiful. Aesthetics cannot be sensed; it can only result from the mind interpretation of object properties.

This thesis focuses on aesthetics as "*pleasure regarded as the quality of a thing*" (Santayana, 1955) and does not consider the developmental, ethical or socio-cultural phenomena associated with aesthetics. Furthermore, the thesis only focuses on the pleasing appearance of *visual* stimuli and does not consider the aesthetics of music, language, mathematical proofs and other non-visual phenomena. Lastly, the thesis focuses only on the visual stimuli that qualify as GUIs or parts of GUIs, and does not consider the stimuli unrelated to HCI, such the patterns of dots and lines from psychology studies or paintings from visual arts.

1.1.3 Methodology

A model of GUI aesthetics would rely on a set of design dimensions to operationalize and estimate aesthetics. The design dimensions could have been outlined and validated using several methodological approaches, which the thesis identifies as the empirical approach, black-box approach, and theory-led approach. The empirical approach (e.g., Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010) relies on user studies and human input. A researcher extracts a large preliminary set of design dimension by analyzing the accounts of GUI aesthetics from artists, designers or regular users. The dimensions may then be operationalized as semantic-differential items and tested in user

studies; the dimensions that do not satisfy psychometric criteria (e.g., confuse participants or do not correlate with similar items) are excluded from the initial set. The exclusion results in the final set of design dimensions, which could be used to evaluate GUI aesthetics. However, because the dimensions are operationalized as semantic-differential items, the evaluation always requires the user input on each dimension for each GUI. Collecting such user input takes substantial time and effort, and is difficult to replicate on a large scale.

Unlike the empirical approach, the black-box approach relies on computation instead of user input, and thus, can be applied on a large scale. The approach implies a researcher initially selects a large number of low-level design features, which might or might not be related to design aesthetics, since the selection is rarely based on a theory (e.g., Ivory & Hearst, 2002; Wu et al., 2011). Such features often cannot be easily attributed to a human-meaningful design dimension, as they describe very low-level visual aspects of design. The large number of initial features also conditions the researcher to use complex statistical methods (such as, support vector machines), which do not explicate the nature of feature-aesthetics relationship. Not knowing the nature of feature-aesthetics relationship – as it could be expressed by a linear, quadratic or another higher-degree polynomial – determines the naming of approach as black-box: researcher sees the input and output of black-box systems, but not the relationship between the input and output. Not knowing the nature of the relationship also makes difficult, if not impossible, the interpretation of features into design improvements. In addition, the absence of feature-backing theory does not let researchers claim that the feature-aesthetics relationship is causal or generalize the relationship to the types of GUIs not tested in researchers' studies.

Similar to the black-box approach, the theory-led approach uses computation to estimate the visual dimensions of designs, but relies on a theory to back the selection of dimensions. The reliance on the theory increases the validity of dimensions and explicates the nature of dimension-aesthetics relationship (e.g., a causal, linear relationship), and thus, eases the interpretation of dimensions into design improvements. The reliance on computation facilitates the prediction of GUI aesthetics on a large scale, as no human input is involved in the prediction once a model is built and validated. This thesis used the theory-led methodological approach and selected the processing-fluency theory to back the outline of design dimension. The processing-fluency theory asserted complexity to underlie aesthetics, which resulted in the outlined dimensions being visual complexity-based.

To validate the design dimensions and their measures (a dimension could have several associated measures), the thesis relied on two groups of methods. The first group of methods involved no human input. The design dimensions were visualized and reviewed for various stimuli (such as, webpages and mobile app layouts). If the visualizations showed a dimension to perform as expected by the processing-fluency theory, the dimension-underlying algorithm and algorithm parameters were retained. Then, if the scores of several dimensions correlated too strongly, the dimensions were merged or dropped – this ensured that each dimension described a unique aspect of design. Finally, dimension measures were analyzed: the strong cross-correlations amongst the measures of the same dimension suggested the high convergent validity of measures, whereas the absence of strong cross-correlations amongst the measures of different design dimensions suggested the high divergent validity of measures.

The second group of validation methods relied on human input: participants rated the visual complexity and aesthetics of various GUIs; the human scores were correlated against computed scores. The significance of correlations confirmed that a dimension indeed described GUI complexity or aesthetics; the direction of correlation – if matched the theory-suggested direction – supported the validity of dimensions. For example, more contour congestion should lead to higher complexity and lower aesthetics, which indeed was the case and was considered a sign of validity of the contour-congestion measure. Further, the scores for several types of GUIs (e.g., for webpages and mobile apps layouts) were collected in the studies. A design dimension should bear a similar effect on all types of GUIs: the high similarity in the reported-computed correlations for different GUIs suggested a higher dimension validity and robustness.

1.1.4 Incremental Development

The analysis of the processing-fluency theory and relevant HCI research resulted in five theoretical assumptions that the thesis work began from. First, visual complexity determines visual aesthetics, as the processing-fluency theory asserts. Second, users can form opinions about aesthetics very quickly

(in under 500 ms, Tractinsky et al., 2006; Lindgaard et al., 2006; 2011). Third, such very-quick opinions are primarily perception based: they depend on the visual dimensions of stimuli since the users have little to no time for cognitive elaboration. Forth, visual dimensions of stimuli determine the visual complexity of stimuli (cf., Reber et al., 2004). Fifth, the dimensions can be quantified automatically. These theoretical assumptions have undergone incremental development and resulted in the predictive model of GUI aesthetics.

The work described in this thesis consisted of two phases: development of GUI aesthetics model, and application of different aspects of the model. The phases are detailed in Chapters 2 and 3 and briefly outlined in Figure 1. Chapters 2 & 3 are based on three published papers each.

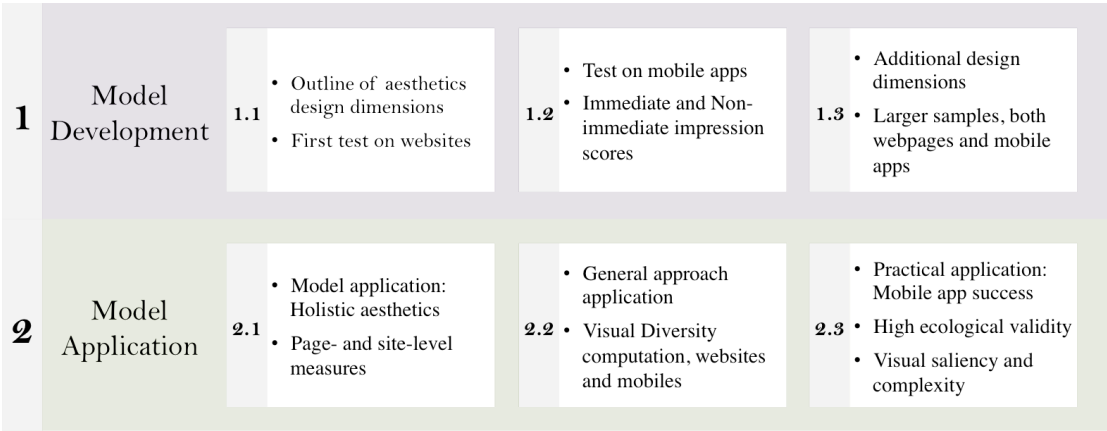


Figure 1. Outline of research undertaken in the thesis.

1.1.4.1 Model Development

A processing-fluency based aesthetics model would rely on several visual complexity dimensions, which needed to be defined and operationalized. Psychology (Oliva et al., 2004; Reber et al., 1999) and HCI (Nadkarni & Gupta, 2007; Forsythe, 2009) had studied visual complexity and described its possible dimensions. However, these dimensions were not outlined in a single source and not operationalized for GUIs in a machine-suitable form – humans could easily understand the descriptions, such as "quantity of detail or color" (Oliva et al., 2004), whereas machines could not and required much more detailed and formalized descriptions.

Chapter 2.1 describes the initial work of collecting, detailing and formalizing the complexity-based design dimensions, and validation of dimensions on the user scores of design complexity and aesthetics. The work was also presented at ACM AVI 2014 (Miniukovich & De Angeli, 2014a). The work began from surveying psychological and HCI literature for the candidate dimensions of visual complexity. The survey resulted in three groups of possible complexity dimensions. First, the perceived *amount of information* on a screen included visual clutter and color variability. Second, the quality of *information organization* included symmetry, grid quality, ease of grouping, and prototypicality. Last, the effort of *information discriminability* (discriminating a symbol from background requires a mental effort, and thus, increases perceived complexity) included figure-ground contrast and contour congestion. Five dimensions were then operationalized and implemented, namely visual clutter, color variability, contour congestion, figure-ground contrast, and symmetry. For some dimensions, existing algorithms were re-used (namely, for clutter and colorfulness); whereas, for the rest, only verbal descriptions were found in the psychological studies – often, studies on the sets of polygons rather than highly-complex stimuli like webpages – and new algorithms had to be created. The dimension measure scores were then computed and matched against the user scores of immediate (50ms exposure) complexity and aesthetics, collected in a user study. The results showed computed scores to correlate with reported scores with a notable exception of color range and symmetry: color range correlated with aesthetics, but not with complexity; and symmetry correlated with complexity, but not with aesthetics. The results also highlighted that the dimension of colorfulness is composed of two sub-dimensions: number of dominant colors (the prominent colors that a human eye can easily distinguish and count) and color range (the whole variety of hard-to-notice hues and semitones that come along with the dominant

colors). The number of dominant colors impacted complexity directly, and aesthetics indirectly; color range impacted only aesthetics. Combining these two sub-dimensions in a single variable of colorfulness might have caused the correlation between aesthetics and colorfulness to be weak in previous research (Reinecke et al., 2013), as the two sub-dimensions correlated with aesthetics in opposite directions. Overall, the automatic measures explained up to 51% of variance in the aesthetics scores of webpages, which was an encouraging result for computational aesthetics.

The focus of the thesis work then shifted towards the evaluation of reliability and ecological validity of the findings. Participants in the initial studies (Chapter 2.1) saw only webpages and only for the very brief 50ms exposure time. Such experimental setup maximized the chance of positive result: 50ms exposure durations allowed neither reading nor conscious elaboration, and participants set ratings relying on low-level visual features (Lindgaard et al., 2011), which the automatic measures were specifically tailored to estimate. The experimental set-up could not test if the measures predicted well the aesthetics of non-webpage stimuli or non-immediate aesthetics. Addressing these concerns, chapter 2.2 describes two studies, which were also presented at ACM NordiCHI'14 (Miniukovich & De Angeli, 2014b). One study extended to mobile apps the claim that immediate aesthetics ($t < 0.5$ sec) carried over into deliberate aesthetics (several sec, Lindgaard et al., 2006; Tractinsky et al., 2006). Such extension suggested that the automatic measures could predict both immediate and deliberate aesthetics of mobile apps. The other study directly tested if the measures could predict mobile app aesthetics (Android apps). The results mirrored the results of sub-chapter 2.1 on webpages: all computed scores correlated with reported aesthetics and complexity scores, with the exception of color range (correlated only with aesthetics) and symmetry (correlated only with complexity). An additional exception was the measure of dominant colors, which no longer correlated with aesthetics. A regression analysis showed the automatic measures explained up to 40% of variance in mobile app complexity and 36% of variance in mobile app aesthetics – less than for webpages, but nonetheless an encouraging result given this was one of the first such studies on mobiles.

The positive results of studies of chapters 2.1 & 2.2 required consolidation, which were targeted in two further studies, one on websites and one on mobile apps. The studies – described in chapter 2.3 and also presented at ACM CHI'15 (Miniukovich & De Angeli, 2015a) – featured two additional automatic measures, more participants, more stimuli, and more-objective sampling techniques. The new automatic measures described grid quality and amount of white space. Also the measure of symmetry was improved, making it more intuitive when visualized. The studies paid close attention to stimuli sampling: biases may be introduced if only few individuals – often, researchers themselves – select the stimuli. The studies instead resorted to crowdsourcing (study 1) or random selection (study 2) as a remedy for such biases: crowdworkers suggested websites from one of three (eCommerce, corporate, news) genres, no more than five websites per worker; an automatic website analyzer randomly picked some apps from Apple's Play Store, again in one of three genres (business, travel and entertainment). The inverse correlation between visual complexity and aesthetics seemed well-established, and participants only rated aesthetics in the studies. The results again corroborated the results of past studies. The automatic-measure scores correlated with user scores as expected; when combined in a linear regression model, the measures explained up to 49% of variance in website aesthetics and up to 36% of variance in iPhone app aesthetics.

1.1.4.2 Model Application

The second phase of the work investigated the application of different aspects of the aesthetics model. The model and computational methods were applied in diverse contexts, such as the estimation of holistic aesthetics, exploration of visual diversity of GUIs, and prediction of mobile app success.

Chapter 3.1 explores the construct of holistic aesthetics, which describes user reaction to "live" websites and apps, not separate static webpages or screenshots that were studied in the first phase of the thesis. Accounting for live-GUI, holistic aesthetics would further increase the accuracy and validity of the predictive model. Several aspects of moving on from static-page aesthetics to the holistic aesthetics were explored, such as the aggregation of individual page scores in a holistic score, the impact of website-level parameters on holistic aesthetics, and importance of below-top-screen parts of pages on the holistic aesthetics. A study (submitted to ACM TOCHI) involved participants browsing and performing information-retrieval tasks on 30 websites, and rating them before and after use. Contrary to the discussions suggesting to include multiple pages in website aesthetics analyses (van Schaik & Ling, 2009), the study showed different-page scores to intercorrelate strongly,

which suggested that one page could stand in for the whole website. A subsequent analysis suggested homepages to suit aesthetics computation better than other pages. Then, the study explored four website-level parameters that could have impacted user impression (number of pages per website, average page length, number of links per page, and average page loading time) and found only the number of pages per website to correlate with user aesthetics scores. Third, the below-top-screen parts of webpages did impact the user, however, the gain in aesthetics-prediction accuracy was small. The top-screen webpage parts could be used for aesthetics estimation. Overall, the study demonstrated that the automatic measures could account for the holistic aesthetics: computed scores did correspond to the reported, after-use aesthetics scores.

Chapter 3.2 applies the computation-based approach to predict GUI visual diversity. Diversity is considered a part of GUI aesthetics (cf., Moshagen & Thielsch, 2010), and only emerges in the live-GUI context, when the user navigates from one GUI layout to another. Visual diversity describes the heterogeneity across different layouts of GUI and can be either harmful – if the user feels lost – or beneficial – if the user feels entertained. The harmful effect could be expected from the processing-fluency principle "the simpler, the better". However, the results of three studies of chapter 3.2 (also ACM BritishHCI'15, Miniukovich & De Angeli, 2015b) showed the latter to be true, at least for the GUIs that were studied. Higher visual diversity led to positive outcomes, suggesting that visual diversity overwhelming the user might be the concern of the past (e.g., such negative effect was described in Grudin, 1989). The only exception emerged in the analysis of shopping mobile apps; popularity correlated inversely with diversity. Such exception might originate from users' dislike for flashy ads, which feature in shopping apps and increase visual diversity. All other indicators of system success – popularity of news and business apps, website aesthetics score, and user view of website-owning companies – correlated positively with diversity. The positive diversity-aesthetics correlation also corroborated the inclusion of diversity in the model of GUI aesthetics (Moshagen & Thielsch, 2010).

Chapter 3.3 (accepted to ACM CHI'16, Miniukovich & De Angeli, 2016) applies the complexity-based visual dimensions from the aesthetics model to study the popularity of mobile apps on Google Play. The study hypothesized that the saliency and complexity of app icons influenced app success: if the user noticed (saliency) and liked an icon (complexity) in a very long Google Play list, they would choose that icon app for a try. In such circumstances of discretionary use, the appearance – which converts into impression very quickly upon a user-system encounter – would be crucial to the success of systems in discretionary-use contexts, as the user spends only short time considering a system and swiftly switches to alternatives if not immediately pleased (cf., Kim & Fesenmaier, 2008). The study estimated the complexity dimensions that were applicable to app icons and one dimension of visual saliency. Results supported the hypothesis. The number of app comments (which strongly correlated with the number of app installs) was taken as the metric of success. The estimates of saliency and complexity correlated with the metric and explained up to 38% of metric variance. Considering that numerous factors could influence app popularity – ranging from the reviews of past users to the number of bugs in an app – explaining 38% of app success by analyzing app icons appeared an encouraging result. The study of chapter 3.3 concludes the work of this thesis, and also provides insights in the future of predictive models of GUI quality: the study features several visualizations of design dimensions, which designers may use for the incremental improvement of designs (Rosenholtz et al., 2011).

1.1.5 Thesis Contribution

The thesis describes the development and validation of the predictive model of GUI aesthetics. The developmental work has produced four major contributions: two theoretical, one practical and one methodological contribution.

The first theoretical contribution adds to the knowledge on aesthetics in HCI by outlining and formalizing several theory-led dimensions of visual design. The studies of chapters 3.1 to 3.3 demonstrate each dimension to describe a unique aspect of design and to correlate with aesthetics linearly (except symmetry). Thus, the studies explicate and detail the links between the dimensions and design aesthetics. Understanding such links is a step forward from vague user's descriptions of designs (e.g., "*the site appears patchy*" or "*the layout is easy to grasp*", Moshagen & Thielsch, 2010) towards identifying specific *quantifiable* design dimensions that made the design look unappealing. Knowing how to quantify such dimensions and how the dimensions relate to aesthetics may suggest

possible improvements to designers.

As the second theoretical contribution, the thesis validates a number of theories and empirical results. First, the thesis provides some evidence that the observed weakness in colorfulness-aesthetics connection (Reinecke et al., 2013) might stem from not distinguishing two aspects of colorfulness, the number of dominant colors and color range (Chapters 2.1 and 2.2.). These two aspects might correlate with aesthetics in opposite directions. Second, the thesis confirms the strong negative correlation between visual complexity and aesthetics, and thus, computationally supports the predictions of the processing-fluency theory (Chapters 2.1 & 2.2). Third, the thesis demonstrates the visual diversity of GUIs to correlate positively with aesthetics (Chapter 3.2). Such observation rejects the hypothesis of visual diversity overwhelming the user (Grudin, 1989) and supports including diversity in aesthetics taxonomies (Moshagen & Thielsch, 2010). Forth, the thesis confirms the link between visual aesthetics and system success (Chapter 3.3). The confirmed aesthetics-success link supports existing theoretical models of system quality (e.g., Hartmann et al., 2008). Last, the thesis confirms a strong correlation amongst immediate (150ms exposure), deliberate (4sec exposure) and post-use (unlimited time, actual use) impressions (Chapters 2 and 3). Such correlation corroborates existing evidence (Thielsch et al., 2013) and suggests that even the shortest, 150ms-based impressions could be the valid measures of aesthetics.

The practical contribution complements the first theoretical contribution: this thesis presents the automatic measures of outlined design dimensions. Such measures have been repeatedly tested on several types of GUIs and could be used in HCI studies as a replacement or addition for the user evaluation of GUI complexity, colorfulness and diversity. The measures could also be used in practice – to aid the evaluation and comparison of designs without user studies, and thus, speed up GUI development. The measure output may further be developed in visualizations, which have also been shown to facilitate design discussions (Rosenholtz et al., 2011).

The methodological contribution includes demonstrating the application of the theory-led approach to model development. Such approach relies on a theory to guide the operationalization of computational features, which results in more valid and robust features than the black-box approach. The black-box approach includes sifting through multiple features in hope that some of them happen to predict the model outcome – such features need to be re-validated on each new type of stimuli.. The theory-led approach also minimizes the amount of user input necessary for model development and validation, relative to the empirical approach. Only the user scores for model output are collected (e.g., the scores of design aesthetics), whereas the empirical approach implies collecting user input for model output and each sub-dimension of model (Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010; Kim et al., 2003; Oliva et al., 2004).

2 MODEL DEVELOPMENT

2.1 Quantification of Interface Visual Complexity

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.

2.1.1 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 (Reber et al., 2004) and HCI (Tuch et al., 2012). Simpler visual stimuli tend to be perceived as more aesthetically pleasing, probably due to more fluent mental processing of stimuli (Winkielman et al., 2002).

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., Sutcliffe, 2009) 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. (Armstrong & Detweiler-Bedel, 2008)), 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 (Reber, 2012) 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.

2.1.2 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 (Reber et al., 2004; Winkielman et al., 2002), which partially persist over time (Tractinsky et al., 2006). The initial judgments further evolve according to conscious elaboration, which could be a source of more intense aesthetic pleasure relative to processing fluency (Armstrong & Detweiler-Bedel, 2008). 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. (Reber et al., 2004)), 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. (2001) 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 (Harper et al., 2009). Depending on the website category, they could explain from 11% to 56% of webpage rating variance. Reinecke et al. (Reinecke et al., 2013) 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. (Wu et al., 2011) 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. (Purchase et al., 2012) 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 (Purchase et al., 2012)) or even simply assumed it to correspond to the number of interface elements (e.g., Ivory et al., 2001). Psychologists, on the other hand, did approach complexity more systematically, distinguishing between amount of information and organization of information (see also, van der Helm, 2000). 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. (2004) 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 2 shows the classification of visual complexity in three main determinants: amount, organization and discriminability of information.

2.1.2.1 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' (Rosenholtz et al., 2007; van den Berg et al., 2009). The variation in color, being the most prominent feature of scene, was also studied separately from visual clutter under the name 'colorfulness' (Reinecke et al., 2013).

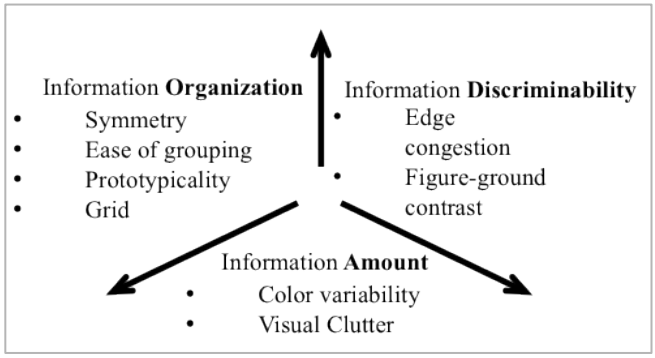


Figure 2. The classification of visual complexity determinants.

2.1.2.1.1 Visual clutter

"Clutter is the state in which excess items, or their representation or organization, lead to a degradation of performance at some task" (Rosenholtz et al., 2007, p.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 (Bravo & Farid, 2004).

Researchers measured visual clutter in several ways. Mack et al. (as cited in, Rosenholtz et al., 2007) suggested using the ratio of edge pixels to all pixels of an image (the Edge Density measure). Rosenholtz et al. (2007) 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 (Tuch et al., 2009). Each of these measures was found to correlate with search performance.

2.1.2.1.2 Color variability

Color variability is a very salient feature of scene. In a study (Oliva et al., 2004), the number of colors ranked second among the most important factors of visual complexity (after the number of objects). However, Oliva et al. (2004) 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. (2003) 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. (2013) 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 (Hasler & Suesstrunk, 2003). They could explain 78% of variation in colorfulness ratings. Similarly, Wu et al. (2011) used the average values and variation in hue, brightness and saturation, and again, colorfulness (Hasler & Suesstrunk, 2003). They could explain 46% of variation in webpage visual quality ratings. Lastly, Purchase et al. (2012) 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 (2012) image colors after reduction), perceived color depth (e.g., Wu et al.'s (2011) variation in hue, brightness and saturation) and idiosyncratic color preferences (e.g., Reinecke et al.'s (2013) preferences for W3C-defined colors). The last category is strongly influenced by acquired tastes (Cyr et al., 2010), and, should therefore be avoided when trying to generalize findings across different demographics.

2.1.2.2 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.

2.1.2.2.1 Symmetry

One of the Gestalt principles, mirror symmetry – the similarity of an object reflection across a straight axis – was claimed to improve interface design (Smith-Gratto & Fisher, 1998-99). 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. (2006) and Tuch et al. (2010) 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. (2009) 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. (2013) 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 (Kootstra et al., 2011; Loy & Eklundh, 2006).

2.1.2.2.2 Ease of grouping

Gestalt psychologists discussed in detail what is perceived as a group and how to increase the ease of grouping (Wertheimer, 1938). Gestalt-based design guidelines (Smith-Gratto & Fisher, 1998-99) 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 (Treisman, 1982) and described several necessary criteria of grouping, such as common location, motion, color, surface, texture, size, and shape (Duncan, 1984; Duncan & Humphreys, 1989). 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.

2.1.2.2.3 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 (Wertheimer, 1938), 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 (Hekkert et al., 2003), 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 (Whittlesea & Williams, 1998). However, despite its significant role in user perception, we found no description of automatic prototypicality measurements. The work on large-scale design mining (Kumar et al., 2013) is a potential step in this direction. It allows studying existing design practices and might serve as a basis for future prototypicality measurements.

2.1.2.2.4 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 (Pothos & Ward, 2000). Only a few studies, though, exploited it in studying interface structure. Reinecke et al. (Reinecke et al., 2013) used quadtree-based symmetry, balance and equilibrium, but not repetition, and reported no influence of these factors on visual complexity. Wu et al. (2011) 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. (2013) 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 Harrington et al., 2004, for a short description of these factors).

2.1.2.3 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.

2.1.2.3.1 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. (2003) 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 (Levi, 2008). For example, the crowding model of visual clutter (van den Berg et al., 2009) 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.

2.1.2.3.2 Figure-ground contrast

Psychologists often use luminance or color contrast to manipulate perceptual fluency. For example, Reber et al. (1999) 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. (2003) 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. (2004) 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.

2.1.3 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. Lindgaard et al., 2011). Finally, we computed six automatic screenshot-based metrics and compared them against user ratings. Visual complexity determinants (Figure 2) that are not explored in the paper were left for future work.

2.1.3.1 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 webpages 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

[1] http://www.smashingmagazine.com/

unattractive/attractive the webpages were" (cf., (Lindgaard et al., 2006; Tuch et al., 2010; 2009; 2012), without any further explanations of the constructs (cf., (Oliva et al., 2004)). After participants signed the consent forms and adjusted their seat height and screen angle, they rated 140 (cf., Lindgaard et al., 2011; 2006; Tractinsky et al., 2006; Tuch et al., 2012) 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., Lindgaard et al., 2011; 2006; Tuch et al., 2012). Third, a black-white noise mask was flashed for 50msec (cf., Tuch et al., 2012) 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., Tractinsky et al., 2006).

2.1.3.2 Automatic metrics

Out of eight identified determinants of visual complexity (Figure 2), 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.

2.1.3.2.1 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 (Rosenholtz et al., 2007)). Edges were detected with the Canny method; low and high thresholds were set to 0.11 and 0.27 (cf., Rosenholtz et al., 2007); 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, (Rosenholtz et al., 2007)). Then, we took file sizes of screenshots saved in JPEG (CL4, the JPEG quality setting set at 70, cf., Tuch et al., 2009). 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.

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	.22					.97	.60	.49	.33

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

[2] http://www.psychopy.org/

[3] We used authors' implementation of Subband Entropy and Feature Congestion (Rosenholtz et al., 2007) cfr. http://dspace.mit.edu/handle/1721.1/37593

Literature describes many measures of color variability (Reinecke et al., 2013; Wu et al., 2011; Purchase et al., 2012; Hasler & Suesstrunk, 2003). We did not consider those that might involve subjective color preferences (e.g., Reinecke et al., 2013) 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 χ^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 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.

#	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. 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 (Hasler & Suesstrunk, 2003)	The measure of colorfulness from (Hasler & Suesstrunk, 2003). It combines standard deviations and means of pixel values taken in the Lab color space.
C9	Luminance SD	The standard deviation of pixel luminance.

Table 2. Implemented measures of color variability.

2.1.3.2.2 Organization of information

As regards the organization of information, we modeled mirror symmetry. The existing methods for symmetry detection (Loy & Eklundh, 2006) 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., Tuch et al., 2010). 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., Rosenholtz et al., 2007) 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 3. 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

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

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:

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

2.1.3.2.3 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.

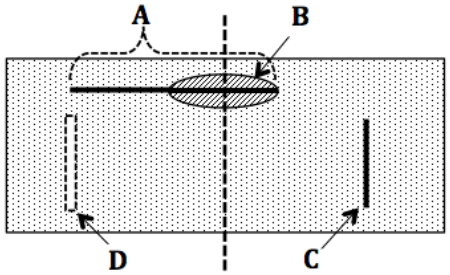


Figure 3. An asymmetrical horizontal line (A) can give symmetrical key points (B); whereas a vertical line (C) cannot, unless it matches another line (D) on the other side.

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

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

2.1.3.3 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.

2.1.4 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 (Reinecke et al., 2013).

	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 3. Pearson correlation coefficients

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 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).

	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; [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)

The results of our user study were in line with similar work (Reinecke et al., 2013; Wu et al., 2011): 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 (Tuch et al., 2012).

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., Reinecke et al., 2013) and suggested that participants might not consider symmetry while

judging aesthetics or that other effects (e.g., demographic, see Tuch et al., 2010) 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 (Tuch et al., 2009; Purchase et al., 2012). Whereas the JPEG file sizes described visual clutter (cf., Rosenholtz et al., 2007), 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 2): GUI prototypicality, grid and easiness of grouping. In addition, the proposed dimensionality of visual complexity (Figure 2) 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.

2.1.5 Conclusion

In this paper, we outlined eight low-level determinants of GUI complexity (Figure 2). 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.

2.2 VISUAL IMPRESSIONS OF MOBILE APP INTERFACES

First impressions are formed very fast but they last. Consecutive approach-avoidance behavior is formed almost instantly and persists over time. The effect of the first impression of graphical user interfaces (GUIs) of desktop webpages on subsequent evaluation is well documented in the literature. Less research has focused on mobile interfaces. To cover this gap, this paper reports two studies. The first study confirmed the persistence of first impressions on mobile interfaces evaluation, although it suggested that exposure time may be longer. The second study extends previous work on automatic evaluation from desktop to mobile interfaces. The linking theme between the studies is that of visual complexity, which is a more objective, yet powerful, predictor of aesthetic evaluation. Using six automatic metrics (color depth, dominant colors, visual clutter, symmetry, figure-ground contrast and edge congestion), in study 2 we explained 40% of variation in subjective complexity scores and 36% of variation in aesthetics scores.

2.2.1 INTRODUCTION

Humans evolved to be very fast at assessing, recognizing and interpreting visual stimuli. GUIs are visual stimuli; they induce an immediate impression (Lindgaard et al., 2006), which largely defines upcoming time-persistent evaluative judgments (Tractinsky et al., 2006). These judgments often determine if a system is going to be used or ignored in favor of rival systems (Tractinsky, 2013). Given the intense competition in the information system domain, first impressions matter to interaction design.

Although there is a relatively new, as yet vibrant, corpus of literature investigating first impressions on desktop interfaces (Lindgaard et al., 2006) relatively little research has addressed mobile interfaces. Researchers (Lindgaard et al., 2006; Bauerly & Liu, 2006; Tractinsky et al., 2006; Tuch et al., 2009; 2012; Zheng et al., 2009) mainly concentrated on desktop versions of websites and their various interaction qualities, including aesthetics. Positive initial impressions are thought to originate from high processing fluency (Reber et al., 2004; Reber, 2012): people like simple interfaces that they understand quickly (Tuch et al., 2009; Miniukovich & De Angeli, 2014a). Smartphone users might value simplicity even more as the on-the-go usage implies multiple external distractions, and short and intensive interaction periods (Choi & Lee, 2012). Quality criteria related to aesthetics are also likely to gain importance in devices people always take with them, through varied social settings.

We conducted two studies on mobile app GUIs. The first study compared evaluations of complexity and aesthetics after a 50ms exposure and 4s exposure. The results agreed with the previous studies on desktop GUIs (Tractinsky et al., 2006; Lindgaard et al., 2006): people do form very fast and stable impression of GUIs. We also confirmed that visual complexity scores correlate with aesthetics scores in mobile design. In the second study, we calculated six automatic metrics of visual complexity for 99 mobile interfaces, and compared them against the subjective evaluations of visual complexity and aesthetics collected from 20 people. The best-fit regression models explained 40% of subjective complexity evaluation and 36% of aesthetics evaluation. These two studies showed how app GUI screenshots could be used for automatic GUI evaluation and extended our initial effort (Miniukovich & De Angeli, 2014a) towards the development of an automatic GUI evaluation tool. In the rest of the paper, we review related work on visual complexity, the measures of visual complexity and specificity of mobile design. Then, we describe study 1, which evaluated the immediate impression of mobiles, and study 2, which quantified perceived complexity and aesthetics of mobile apps. Lastly, we summarize the results.

2.2.2 Related work

Lindgaard et al. (2006; 2011) established that users form reliable first impressions of desktop webpages almost instantly, in the first 50 ms. Tractinsky et al. (2006) further confirmed these impressions persist over longer time: the ratings of desktop webpage aesthetics after a 500 ms

exposure correlated well with the ratings after a 10 s exposure. Since then, researchers (Zheng et al., 2009; Tuch et al., 2009; Miniukovich & De Angeli, 2014a) regularly used very short exposures for studying visual complexity and aesthetics of desktop GUIs.

Psychology (Reber et al., 1999; 2004; Reber, 2012) offers an explanation of the origins of immediate impression: the fluency with which we process stimuli serves as a shortcut for deciding if a stimulus is unknown, possibly dangerous and should be avoided, or familiar and bears no danger. Processing fluency affects our subjective judgments of familiarity (Whittlesea & Williams, 1998), simplicity and aesthetics of stimuli (Reber, 2012). More advanced, conscious considerations may also affect these judgments too, and possibly, to a larger extent. However, applying conscious considerations in user studies stays problematic due to a huge variety of idiosyncrasies an individual may have. Processing fluency, on the other hand, is relatively universal across all cultures and social groups (Reber, 2012).

Partially based on the notion of processing fluency, Choi et al. (Choi & Lee, 2012) proposed and validated a structural model of simplicity for smartphone interfaces. They approached interfaces not as purely visual things, but as a combination of visual, cognitive and functional aspects. The model consisted of three dimensions: visual aesthetics, information design and task complexity. The visual aesthetics dimension stemmed from Lavie et al.'s (2004) *classical aesthetics* and Moshagen et al.'s (2010) *simplicity*, and accounted for visual configuration factors: clarity, orderliness, homogeneity, grouping, balance and symmetry. The information design dimension stemmed from the need to reduce cognitive load and to structure the information around us, i.e., the information should be reduced, organized, integrated and prioritized. The last dimension, task complexity, stemmed from the interaction-oriented nature of interfaces: all interfaces afford performing certain actions. The interdependence between actions, the clarity of action outcome and visual information cues designating the actions were the sub-dimensions of task complexity. In this paper, we solely concentrate on visual complexity, the concept corresponding to Choi et al.'s (2012) visual aesthetics dimension.

2.2.2.1 Visual Complexity

Unlike the proponents of the "*unity in diversity*" principle (e.g., Moshagen & Thielsch, 2010; Hekkert et al., 2003), we did not differentiate conceptually between simplicity and complexity (also, diversity), but viewed them as the opposite sides of the same dimension. In this, we continued our earlier work on GUI visual complexity (Miniukovich & De Angeli, 2014a), where we considered low-level, preconscious determinants of visual complexity. We (Miniukovich & De Angeli, 2014a) listed and classified these determinants in three groups: the amount of information, the organization of information and discriminability of information (Table 6). While the amount and organization already existed in theories of information complexity (see Donderi, 2006, for a review), we specifically added discriminability to account for *visual* in visual complexity.

2.2.2.1.1 Amount of Information

The notion of information amount was formalized in theories of arithmetic and probabilistic complexity as the size of minimal-length data needed to restore an original object (Donderi, 2006). Psychologists adopted this view and argued the minimal length corresponds to the minimal work for the brain, and therefore, to energy economy (Aksentijevic & Gibson, 2012). From the physiological viewpoint, the diversity of information corresponds to the range of stimulation input a human can perceive. In the visual domain, this means variation in color, luminance, orientation, form and motion. HCI researchers studied this phenomenon as 'visual clutter' and measured it with edge density (Rosenholtz et al., 2007), feature congestion (Rosenholtz et al., 2007), subband entropy (Rosenholtz et al., 2007) and JPEG file size (e.g., Tuch et al., 2009) measures.

When asked to elaborate why some stimuli seem complex and others do not (see Oliva et al., 2004), participants mention color variability much more often than other low-level scene aspects (e.g., figure-ground contrast). Thus, due to its apparent prominence and "eye-catching" potential, the variability in color was often explored independently from visual clutter (cf., Ling et al., 2007). Researchers tried to count the number of colors after color reduction (Purchase et al., 2012), compute the mean (Reinecke et al., 2013; Wu et al., 2011; 2013) and standard deviation (Purchase et al., 2012) of pixel hue, saturation and value, and estimate the relative importance of particular colors (Reinecke et al., 2013). However, a part of these measures was related to idiosyncratic color preferences, which are largely based on nurture, are not universal, and for this reason have been excluded from our

research scope. Our past research (Miniukovich & De Angeli, 2014a) suggested all color-based measures should be further divided in two sub-categories: color depth and dominant colors. Color depth describes very gradual transitions between two colors, and lets designers create the impression of 3D objects on 2D screens, and smooth corners and "saw-like" diagonal lines. Dominant colors describe the number of visually distinct colors on a screen. Too many dominant colors make the user perceive a GUI as complex (Miniukovich & De Angeli, 2014a). However, if combined, these two color metrics explain little of perceived visual appeal (cf. Reinecke et al., 2013).

Information:		
Amount	Organization	Discriminability
• Clutter • Dominant colors • Color depth	• Symmetry • Ease of grouping • Prototypicality • Grid	• Figure-ground contrast • Edge congestion

Table 6. The classification of visual complexity determinants.

2.2.2.1.2 Organization of Information

The probabilistic theories of information complexity (Donderi, 2006) assumes structured information can be compressed better, and therefore, requires less space for storing. Research in psychology and HCI found a preference for symmetrical, easy to group, prototypical and grid-based visual arrangements. Mirror symmetry – a reflection of an object across a straight axis – was mainly studied in perceptual psychology using abstract dot and line arrangements. However, higher levels of symmetry also positively correlated with user preferences for webpage designs (Bauerly & Liu, 2006), and was often used for aesthetic document formatting (Balinsky, 2006). Ease of grouping is based on several visual properties (e.g., color or motion, see (Duncan & Humphreys, 1989) for more detail) shared within members of a group. Clear visual resemblance within group members and clear visual difference from other groups aid visual "parsing" of the scene and are included in interface design guidelines (Smith-Gratto & Fisher, 1998-99). Prototypicality is not an innate but acquired property of human perception. Visual arrangements that are often seen become a standard; they are recognized quicker (Whittlesea & Williams, 1998) and favored more. In the online world, people form relatively stable expectations how a typical webpage should look depending on the website type (Roth et al., 2010). Strong deviations from the standard induce aversion, whereas subtle deviations are seen as novel and interesting (Hekkert et al., 2003). A grid-based structure eases orienting within a visual scene; regularity and repetition contribute to figural goodness (Pothos & Ward, 2000; Harper et al., 2013). Basing a GUI on a grid has become a standard practice and is extensively used in automatic document generation (Balinsky, 2009). However, the quality of grids can also vary across GUIs, resulting in various levels of GUI quality.

2.2.2.1.3 Discriminability of Information

The discriminability of information does not affect the extent to which information can be compressed, but rather the effort of inputting the information into the visual system. Possible difficulties in information inputting follow from the natural limits of human visual system (e.g., the resolution ability of the human eye). Participants often describe complexity as dependent on figure-ground contrast and open space (Oliva et al., 2004), two constructs unrelated to either amount or organization of information. The lack of open space leads to edge congestion, a particularly prominent issue in large-scale graph visualizing, where numerous graph edges obscure other edges and nodes (Wong et al., 2003). The minimal acceptable distance between visual objects is called critical spacing, which, if violated, results in visual input losses (Levi, 2008). Figure-ground contrast describes a luminance or color difference between two adjacent areas. Psychologists often use various levels of contrast to manipulate the speed of phrase reading or object recognition (Reber et al., 1999). Thus, higher contrast should lead to higher processing fluency, and therefore, lower the complexity of stimuli.

2.2.2.2 Measures of Visual Complexity

HCI researchers have measured visual complexity of GUIs using two different approaches. Subjective measures are based on collecting and processing user reactions to seeing or interacting with an interface. Automatic measures are based on extracting and processing features of interfaces or features of screenshots of interfaces.

2.2.2.2.1 Subjective Measures

Several methods have been employed to collect subjective user input. Different variations of card sorting methods have helped to study what complexity means to participants. Oliva et al. (2004) asked participants to split 100 indoor scene images in groups based on scene complexity and to describe the criteria they used for grouping. Participants described complexity as a derivative from the number of scene objects and colors, used open space, symmetry, organization or contrast. Harper et al. (2009) asked participants to sort 20 webpage photographs by their complexity and to elaborate on sorting criteria. Unlike Oliva et al. (2004), the criteria Harper et al. collected were very domain specific and included the number of webpage external links, the amount of text, the sizes of tables, boxes, lists and images and several other website-related parameters.

Multi-item questionnaires, the most often-used input collection method, are suitable for exploring the interrelationships between various aspects of GUIs. Lavie et al. (2004) developed and validated a multi-item scale of classical aesthetics, which included such concepts as clean, clear and symmetrical, and which was later re-used in the context of simplicity (Choi & Lee, 2012). Moshagen et al. (2010) developed another multi-item scale of aesthetics, which included a simplicity dimension. The items of simplicity consisted of perceived layout density, easiness to grasp, consistency and quality of structure. Tuch et al. (2009) concentrated specifically on visual complexity of webpages and measured it with three items: visually complex, well organized and overloaded. The consistency between these items was high (Crobach's alpha = .80), suggesting a single-item measure could be used relatively safely.

Another method of input collection, single-item questionnaires, is less reliable than its multi-item counterpart, but lets participants evaluate more stimuli in same-duration experimental sessions. Researchers often used single-item methods in studies on automatic measures of GUI complexity, where a large number of GUIs secures the representativeness of GUI sample. Wu et al. (2013) asked seven participants to rate the complexity of 1300 webpages in one session, which, however, could be too many, since it resulted in high inter-rater score variability. Reinecke et al. (2013) studied complexity and colorfulness of 450 webpage screenshots. The visitors of their online lab (184 person) rated 30 screenshots each, which was short enough to exclude fatigue effects. However, they had to exclude a significant number of screenshots (68 for complexity) from the analysis due to high inter-rater disagreement, which could, in part, be caused by varying environmental conditions (e.g., the lighting of room or size of screen). In our earlier work (Miniukovich & De Angeli, 2014a), we attempted to minimize such variations and tested all our participants in the same lab. Ten people rated complexity and attractiveness of 140 webpage screenshots. The inter-rater consistency was high (ICC_{comp} = .69; ICC_{attr} = .81).

2.2.2.2.2 Automatic Measures

Asking participant rating is expensive and time-consuming. In response, researchers have proposed several automatic measures of GUI complexity, which could be split in element-based and pixel-based. Element-based metrics consider standard GUI elements, such as buttons, labels and icons. Harper et al. (2013) explored the distribution of visual blocks on webpages and predicted webpage ranking by participants with 86% accuracy. Ivory et al. (2001) described several element-based metrics of webpage quality, including the number of words, font types and sizes, and links per page, the proportion of emphasized, visible and invisible text, number of text colors, and reading complexity. Depending on a website category, they could explain from 11% to 56% of webpage rating variance (Ivory et al., 2001). Lastly, Purchase et al. (2011) applied several aesthetics metrics to 15 webpages. Using rectangular areas of all elements of webpages, they measured webpage balance, symmetry, cohesion, equilibrium and several other metrics. A part of the metrics could predict user ranking of webpage visual appeal.

Pixel-based metrics consider GUIs as sets of pixels and can be applied to any type of GUI without substantial adjustments. Zheng et al. (2009) partitioned 27 webpage screenshots in color-homogeneous blocks and used the distribution of these blocks to assess webpage symmetry, balance and equilibrium. A part of estimates correlated with the ratings of webpage visual appeal. Wu et al.

(2011) also partitioned webpages into blocks, but using webpage underlying HTML structure. They applied 30 different metrics to the blocks, including the ratio of block dimensions, the average block hue, saturation and value, and the number of visual blocks. Their metrics accounted for 43% of variance in user ratings of visual appeal. Reinecke et al. (2013) used a variety of algorithms to describe webpage colorfulness and visual complexity. They calculated the mean of pixel hue, saturation and value, a measure of colorfulness, the number of text and image areas, and the measures from (Zheng et al., 2009). All the measures together explained up to 48% of user ratings of webpage visual appeal. Lastly, Purchase et al. (2012) explored the visual complexity of complex scene images. They quantified it with the number of image colors (before and after color reduction), the standard deviation of pixel luminance, the proportion of contour pixels to all pixels, and file sizes of images saved in JPEG, PNG and GIF formats. The proportion of complexity rating variance they explained was close to 25%.

2.2.2.3 Mobile Design Specificity

Although mobile devices take up more and more functions traditionally done on desktop devices, they still differ from desktops in several respects. Researchers (Chittaro, 2011; Choi & Lee, 2012) generally emphasize three major differences: screen size, context of use and interaction modalities. For portability reasons, the screens of mobiles are small. Multiple overlapping windows cannot be displayed efficiently on them. There is no room for much white space, and every smallest bit of screen space is valuable. This urges mobile GUI designers to pack as many elements on a single screen as possible, and therefore, increases the complexity of mobile interfaces relative to desktop interfaces.

The context of mobile device usage is also very different: on the go, away from power sources and in various lighting conditions. Bohmer et al. (2011) explored usage practices of 4100 Android users: an average user interacted with their mobiles one hour a day, but a typical interaction session lasted less than a minute, with 50% of all sessions shorter than 5 seconds. Combined with less bright screens due to battery life economy and problematic lighting conditions (e.g., while in sunlight), this argues against having high visual complexity. Mobile users do value well-structured visual configurations, shorter lists and simpler menus (Chae & Kim, 2004).

Interacting with a mobile differs from interacting with a desktop (Chittaro, 2011). A variety of mobile-only sensors (e.g., GPS, accelerometers and physiological sensors) still cannot substitute for conventional keyboards and mice. Using fingertips as pointing, scrolling, zooming and selecting devices is much less precise than using mice. Preventing input errors on mobiles requires much bigger interface elements than those for a traditional mouse (cf., Choi & Lee, 2012).

Overall, designers face very tough challenges if they try to account for all aforementioned requirements in a single mobile GUI. Even more challenging, they often need to account for the smartphone's visual appearance. Smartphone appearance explains up to 30% of user satisfaction and is the main consumer choice factor (Ling et al., 2007), which is not surprising, given that people always wear mobiles on them and use it to enhance their personal image.

2.2.3 Study 1

The first study aimed at confirming the results of past studies (Lindgaard et al., 2006; Tractinsky et al., 2006) on another type of stimuli (mobile app GUIs rather than widescreen website GUIs). Past studies have shown that the first impressions of website GUIs are formed very quickly and persist with time. However, this was not tested on mobile app GUIs. Before we tried to quantify the first impressions of mobile app GUIs (see Study 2 below), we needed to ensure such impressions would form quickly and persist (as it is for website GUIs).

2.2.3.1 Method

We based our experimental procedure on a study looking at Web GUIs (Tractinsky et al., 2006, experiment 1), adapting their approach to the specificities of mobile devices.

2.2.3.1.1 Stimuli

We started by sampling mobile GUIs using a set of criteria. First, we only considered mobile apps and not mobile websites. Literally every website could be opened with a mobile browser, and it was often unclear which website was designed specifically for mobile platforms and which was not. We deliberately concentrated on GUIs for mobile platforms. Second, we only considered free apps from

the Travel & Local, Entertainment, Business and Social categories of Google Play. This choice of categories should have ensured a required diversity of apps on the task-fun dimension. Third, we took three screenshots per app. We excluded overly simplistic apps that did not have three visually different screens (e.g., a screen featuring a list of items is visually different from a screen featuring the details of an item, see Figure 4). Fourth, we added content to the apps that looked unnaturally empty (e.g., a messaging app without any messages), but avoided to add and select screens with emotionally charged content (e.g., big images featuring smiling people). Fifth, we avoided selecting screens featuring no interface controls, that is, only plain text or a single photograph. Sixth, we did not select overlays and menus that did not cover the entire screen. Finally, we wanted to reduce possible familiarity biases, and therefore, tried to avoid apps installed more than a million times (Google Play shows the number of installs for each app). After we selected 17 apps and took 51 screenshot (480×762 pixels, without the top bar; PNG format, truecolor, 24-bit per pixel), we validated our selection: nine members of a local HCI research group reviewed the screenshots and commented if the screenshots truly represented the apps. Several screenshots were re-taken based on the comments.

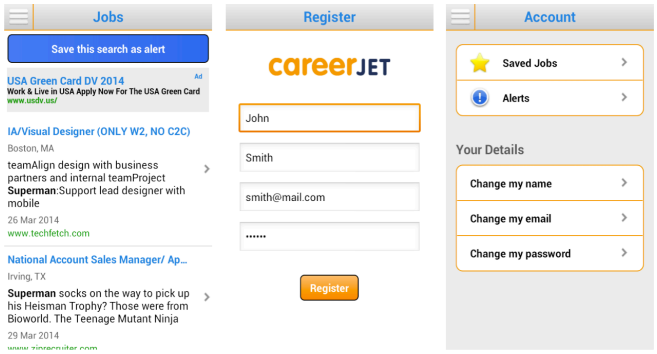


Figure 4. A selection of screenshots of a business app. We tried to take visually diverse screenshots of different app activities.

2.2.3.1.2 Participants

Seventeen students (mean age = 24.9 year; SD = 3.2; 14 – male; all fluent in English) of a CSCW class participated in a group experimental session on a voluntary basis. One participant did not finish the test. The data from three more participants were excluded due to very low (close to zero) correlations between their ratings and the means of ratings of others. Further inspection of data showed these participants did not take the task seriously: their answers contained sequences of same-number ratings (e.g., eight "4" in a row) and their average time-to-rate was well below one second (the time of others was ~1.5 – 2.0 sec).

2.2.3.1.3 Experimental Design

We used a two-way experimental design with the exposure duration as a within-subject independent variable (50ms VS 4s). The duration of 50 ms was used in similar studies (Lindgaard et al., 2006; Miniukovich & De Angeli, 2014a) and deemed to be brief enough to allow very little conscious elaboration. The duration of 4s has been used in studies of complex scene viewing; it enables up to 10 eye fixations (eye fixations average at 300ms (Sereno & Rayner, 2003)) and allows for conscious elaboration. Thus, our experimental manipulations related to the amount of conscious elaboration: 50 ms did not allow for it, 4 s did allow for it.

Each participant rated each of 51 screenshots twice, once after seeing it for 50 ms and once after seeing it for 4 sec. A similar past study (Tractinsky et al., 2006) used an identical procedure for all participants (short exposure first, long exposure second), and the difference between their two conditions might have been attributed to learning and mere-exposure effects. We instead balanced our experimental conditions: half of the participants started with the 50 ms exposure, the others started with the 4 s exposure. Our experimental procedure was implemented as a native Android app. Before the test, we asked participants to set the brightness of their mobile screens to the maximum and disable external notifications. After reading a briefing form, signing a consent form and

providing basic demographic information, participants downloaded and installed the app on their mobiles (Samsung Galaxy SII, for all but two participants), and entered the test.

A trial for each screenshot consisted of several steps. First, participants focused their sight on a red fixation cross in the middle of screen for 1-1.5 sec. Second, a screenshot was shown for 50 ms or 4 sec, depending on the current experimental condition. Third, a black-white noise mask was shown for 50 ms. Finally, participants rated the screenshot on a 7-point semantic differential scale. Each participant rated only one quality of screenshots, either visual complexity (1 – simple; 7 – complex) or visual aesthetics (1 – ugly; 7 – beautiful). Rating both qualities could increase the task difficulty and bias participants (due to a noticeable link between complexity and aesthetics).

2.2.3.2 Results

In total, seven rated visual aesthetics and six rated visual complexity. In the 4 s condition, the inter-rater consistency was high for both aesthetics (ICC = .85; 95% conf. interval .78 < ICC < .90) and complexity (ICC = .75; 95% conf. interval .63 < ICC < .85). In the 50 ms condition, the inter-rater consistency was lower, but still acceptable for both aesthetics (ICC = .54; 95% conf. interval .31 < ICC < .71) and complexity (ICC = .54; 95% conf. interval .31 < ICC < .71). These levels of consistency indicated participants understood the task and key constructs (complexity and aesthetics).

Participants formed stable evaluative judgments very quickly. The between-subject mean ratings in the 50ms condition correlated with the mean ratings in the 4s condition for both aesthetics ($r = .77$, $p < .001$; Figure 5) and complexity ($r = .58$, $p < .001$; Figure 5). The average within-subject correlations were also relatively high (between ratings in the 50ms and 4s conditions) for both aesthetics ($r_{avg} = .48$; from .34 to .75) and complexity ($r_{avg} = .30$; from .13 to .51). Paired t-tests showed no significant difference between the mean ratings in the 50ms and 4s condition for both aesthetics ($t = 1.69$; $p = .10$) and complexity ($t = -.89$; $p = .38$). This remained true even if the t-tests conducted separately for the 50ms-first condition ($t_{attr} = .97$, $p = .34$; $t_{comp} = -.66$, $p = .51$) and 4s-first condition ($t_{attr} = 1.77$, $p = .08$; $t_{comp} = -.82$, $p = .42$). Finally, we also observed a negative correlation between the mean ratings of complexity and aesthetics in both 50ms condition ($r = -.36$; $p < .01$) and 4s condition ($r = -.63$; $p < .001$).

2.2.3.3 Discussion

The results suggested the immediate-impression principles from the Web domain (Lindgaard et al., 2006) also function in the mobile app domain. Participants did not alter their evaluations drastically between the 50ms and 4s conditions. The magnitude of correlations was only slightly lower than in similar studies (Tractinsky et al., 2006, experiment 1), which might have reflected the shorter exposure durations we used. Overall, our results supported the use of immediate impression in HCI studies on mobiles: it formed very quickly and persisted with time, and therefore, affected the general attitude towards mobile products.

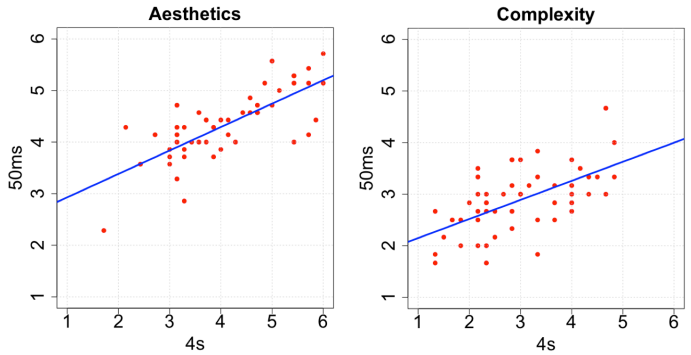


Figure 5. Correlations of 51 screenshot mean ratings of aesthetics (left) and complexity (right) in the 50ms and 4s conditions.

Also consistent with past studies on desktop GUIs (Tuch et al., 2012; Miniukovich & De Angeli, 2014a), our ratings of complexity correlated negatively with the ratings of aesthetics. The

correlations after the "cognition-allowing" 4s exposure were even stronger than after the "perception-only" 50ms exposure. Thus, even if the user impression of mobile GUI quality changed over time due to cognitive elaboration, GUI complexity still stayed largely important for mobile users (as suggested in Choi & Lee, 2012).

However, we decided to not use the user ratings from this study due to an issue: the consistency of 50ms ratings was much lower than the consistency of 4s ratings. This could be due to two reasons. First, the 50ms condition required much concentration and full attention, which was not possible in the group experimental settings. Second, 50ms could be too brief to elicit significant response. We addressed these concerns in study 2.

2.2.4 Study 2

In this study, we tried to accurately collect the ratings of user immediate impression of complexity and aesthetics, and used them as a benchmark, to compare against the scores our automatic metrics (Miniukovich & De Angeli, 2014a) produced.

2.2.4.1 Method

The methods of study 2 resembled those of study 1. However, the primary purpose of study 2 was collecting immediate user impressions. Therefore, we cut out the long, 4s exposure condition.

2.2.4.1.1 Stimuli

A linear regression analysis requires a relatively large sample of data points. Therefore, we enlarged the sample used in the first user study from 51 to 99 screenshots (33 apps). Additional screenshots were selected using the same criteria as in the first study.

2.2.4.1.2 Participants

Twenty doctoral students and postdocs (mean age = 30.9 year, SD = 5.7; 7 – females; all fluent in English; all regular app users) participated in the second user study. All participants reported having experience with apps (the mean number of apps used per day = 4.4; SD = 2.4).

2.2.4.1.3 Experimental Design

The test procedure was the same as in the first study with two exceptions: there was only one experimental condition (short exposure) instead of two (short and long exposures); the duration of the screenshot exposure was 150ms instead of 50ms or 4s. The duration of 150ms is shorter than the 200ms required for reading a word (Sereno & Rayner, 2003), but longer than the 107ms required to reliably grasp the gist of complex scene (Fei-Fei et al., 2007). The study was conducted individually, in the same room with the same experimenter, on the same Samsung Galaxy SII. The brightness of the screen was set to the maximal level; all notifications were disabled.

2.2.4.2 Automatic Metrics

Earlier (Miniukovich & De Angeli, 2014a), we explored the automatic metrics of desktop web GUI complexity and aesthetics. The metrics were classified in three groups (see also Table 6): information amount (dominant colors, color depth and visual clutter metrics), information organization (a symmetry metric) and information discriminability (figure-ground contrast and edge congestion metrics). Only the metrics of information amount consisted of multiple sub-measures, which did not let us test the classification using regular factor analysis. In the present study, after only minor adjustments, we re-applied the metrics (Miniukovich & De Angeli, 2014a) to mobile app GUIs and correlated the outcome scores with the user ratings of mobile GUI complexity and aesthetics.

2.2.4.2.1 Amount of Information

For each mobile app screenshot, we computed several measures of information amount and clustered them into three metrics of information amount using a single 3-factor maximum-likelihood factor analysis with Varimax rotation. The loadings of the measures (A1-A12) are in Table 7. Two clustered-color-based measures loaded on a factor, which we interpreted as the metric of dominant color. The first measure (A1) described the number of static clusters. Pixel RGB values of each screenshot were clustered in 512 same-size clusters (32^3 possible color combinations per cluster) and the number of clusters was counted. Color variability was reduced: the RGB values used less than three times per screenshot were filtered out. We reduced this filtering threshold from five (as in Miniukovich & De Angeli, 2014a) down to three pixels due to the smaller size of mobile screenshots,

and therefore, smaller pool of pixels to choose from. The second measure (A2) described the number of dynamic clusters. Pixel RGB values of each screenshot were first put in a color cube. Then, if, in the color cube, the distance between two colors was less than or equal to three, these colors were united in the same cluster. This process continued recursively for all colors in the color cube. Color variability was also reduced, but we decided to preserve the five-pixel cut-off threshold (as in Miniukovich & De Angeli, 2014a) due to the dynamics of this clustering method: too low threshold would leave "trails" between real color clusters and the clusters would get united.

Three color-based measures loaded primarily on a factor that we interpreted as color depth (Table 7). Following the procedure from (Miniukovich & De Angeli, 2014a), we counted distinct pixel RGB values met more than two times (A3); recorded the file sizes of screenshots in the PNG format (A4); and counted distinct Hue, Saturation and Value pixel values coveting more than 0.1% of screenshot (A5-A7). We also added a new parameter (A8) describing the ratio of unique RGB values to the number of dynamic clusters (aka, A2 measure). Three parameters (A5-A7), however, were excluded from the factor analysis because of the violation of normality requirement of factor analysis.

The metric of visual clutter combined four known measures of visual clutter (Table 7). These measures were the ratio of edge pixels to all pixels (A9), Subband Entropy (A10, Rosenholtz et al., 2007), Feature Congestion (A11, Rosenholtz et al., 2007), and sizes of JPEG-compressed images (A12). All measure algorithms and algorithm settings were identical to those from (Miniukovich & De Angeli, 2014a).

measures / factors	A1	A2	A3	A4	A8	A9	A10	A11	A12
Dominant colors	.54	.98	.44	.28		.19	.16	.20	.19
Color depth	.40	.13	.77	.88	.77	.33	.11	.31	.49
Clutter	.29	.17	.24	.27	.30	.84	.96	.81	.78

Table 7. Factor loadings of information amount measures (cumulative var. g = .83). We used .50 as a cut-off value.

2.2.4.2.2 Organization of Information

The metric of GUI symmetry was based on a single measure, as in (Miniukovich & De Angeli, 2014a). The measure considered contour pixels of screenshots. If a contour pixel had a counterpart across the main vertical axis, it was counted as a symmetrical pixel. Then we took the ratio of symmetrical pixels to all edge pixels and normalized it by edge density (Rosenholtz et al., 2007). The measure settings were identical to those from (Miniukovich & De Angeli, 2014a) with an exception of symmetry tolerance threshold, which was twice as high (the size of Galaxy SII pixels was approximately a half of pixel size of the MacBook we used in (Miniukovich & De Angeli, 2014a)).

2.2.4.2.3 Discriminability of Information

The metric of GUI contrast reflected the salience of contours of GUI screenshots. Every contour (aka, an edge) is the product of luminance or color difference between adjacent pixels. Higher difference means higher contrast. We detected and counted the proportions of edge pixels on several consecutive luminance levels. Then we weighted the proportions (subtler edges received higher weights), summed them up and normalized by the overall number of edge pixels. The received number was the metric of contrast we used. Algorithm settings were identical to those from (Miniukovich & De Angeli, 2014a) with one notable exception: consecutive luminance levels varied from 0.05 to 0.65 (rather than from 0.1 to 0.7). We made this adjustment to account for background texture contours, which seemed to be subtler in mobile GUIs than in desktop GUIs.

Lastly, the metric of edge congestion described the proportion of edges that were too close to other edges. First, we detected edges. Then, if an edge pixel was in a 20-pixel vicinity with a pixel of another edge, we marked it as congested. The ratio of congested pixels to all edge pixels was the congestion measure we used. Algorithm settings were identical to those from (Miniukovich & De Angeli, 2014a). Despite the difference in pixel sizes (and therefore, in the number of pixel per centimeter) between a Galaxy SII screen and MacBook screen, doubling the 20-pixel threshold seemed impractical (40 pixels were 1/12 of the entire Galaxy SII screen width).

2.2.4.3 Results

All participants completed the test; no data were excluded. The inter-subject consistency was high for both aesthetics (ICC = .79; 95% conf. interval .73 < ICC < .85) and complexity (ICC = .81; 95% conf. interval .74 < ICC < .86). The selection of stimuli was balanced, as reflected by the means close to the scale center (sc_{median} = 4) for both aesthetics (mean = 3.92, SD = .67) and complexity (mean = 3.58; SD = .85) and very small skews ($skew_{aest}$ = -.04; $skew_{comp}$ = -.07). The magnitude of negative correlation between the mean ratings of complexity and aesthetics was moderate (r = -.29; p < .01).

The scores of all six metrics correlated with either visual complexity or visual aesthetics or both (Table 8). We tested the mediation role of complexity on the relationships between aesthetics and clutter, contrast, and edge congestion (Table 9). The results of the Sobel tests however suggested the mediation was not significant for all cases: aesthetics and clutter (z = -1.51; p > .05), aesthetics and contrast (z = 1.71; p > .05), and aesthetics and congestion (z = -1.91; p > .05). Thus, there was no significant mediation effect. The best-fit regression model of visual complexity included the metrics of dominant colors, visual clutter, contrast and symmetry, and explained 40% of complexity score variance (Table 10A). The best-fit regression model of visual aesthetics included the metrics of color depth and clutter, and explained 36% of aesthetics score variance (Table 10B). When we added complexity scores in the regression model of aesthetics as an independent variable, the explained variance increased to 38% (Table 10C).

	Aesthetics		Complexity	
	r	**p**	**r**	**p**
Dominant colors	.01	> .05	.29	< .01
Color depth	.52	< .001	-.03	> .05
Visual clutter	-.32	< .01	.51	< .001
Contrast	.35	< .001	-.35	< .001
Edge congestion	-.25	< .05	.45	< .001
Symmetry	.05	> .05	-.44	< .001

Table 8. Pearson correlation coefficients.

Independent variables		β	p	Indirect effect	z*
Clutter	before	-.32	<.01	-.09	-1.51
	after	-.23	<.05		
Figure-ground Contrast	before	.35	<.001	.07	1.71
	after	.28	<.01		
Edge Congestion	before	-.25	<.05	-.10	-1.91
	after	-.15	.16		
* the significance threshold value is 1.96					

Table 9. Mediator analysis (complexity - mediator; aesthetics - DV): β-scores of linear models before and after adding the mediator, and magnitude of indirect effect (Sobel test).

2.2.4.4 Discussion

The inter-subject consistency for both aesthetics and complexity ratings was high. This suggested we succeeded in solving the issues of study 1 and could safely use collected ratings as true measures of immediate perceived visual complexity and aesthetics. Similar to the 50ms condition of study 1, the magnitude of aesthetics-complexity correlation (r = -.29; p < .01) was moderate and lower than an analogous correlation in our past study on desktop GUIs ((Miniukovich & De Angeli, 2014a), r = -.55; p < .001). This difference could reflect an improvement in experimental design: in the present studies, different people evaluated complexity and aesthetics, and therefore, could not carry over their impression of one construct on another.

	Outcome	Predictor	β	p	R^2 (R^2_{adj})
A	Visual complexity	Dominant colors	.27	< .05	.40 (.37)
		Visual clutter	.35	< .001	
		Contrast	-.16	.069	
		Symmetry	-.19	< .05	
B	Visual aesthetics	Color depth	.51	< .001	.36 (.35)
		Visual Clutter	-.31	< .001	
C	Visual aesthetics	Color depth	.58	< .001	.38 (.36)
		Visual clutter	-.22	< .05	
		Visual complexity	-.15	.092	

Table 10. Linear regression models of a) perceived visual complexity, b) visual aesthetics and c) visual aesthetics with complexity included.

We re-used the automatic metrics of GUI complexity as they were described in (Miniukovich & De Angeli, 2014a), with only a few minor adjustments in algorithm settings. In particular, we lowered the color reduction threshold, which followed from the smaller screen size; lowered edge detection thresholds, which followed from subtler background textures; and increased our symmetry tolerance threshold, which followed from the smaller absolute size of mobile pixels. We also clustered all information-amount-related measures in a single maximum-likelihood analysis rather than in two separate analyses (one for clutter, and the other for color depth and dominant color metrics). Such a single-analysis approach reflected better our classification of complexity dimensions (Figure 4). Only minor adjustments in algorithms (c.f., Miniukovich & De Angeli, 2014a) and similar correlation magnitudes (c.f., Miniukovich & De Angeli, 2014a) supported the use of our metrics as reliable indicators of GUI complexity and aesthetics.

Consistent with our past results for desktop interfaces (Miniukovich & De Angeli, 2014a), our metric scores correlated with user scores of complexity and aesthetics. As the only significant difference, dominant colors no longer described aesthetics scores. This might reflect the overall lower number of dominant colors used in mobile GUIs, i.e., "patchiness" was not common and was not an issue in mobile interfaces. Notably, the effect of symmetry on complexity was much stronger. Unlike in desktop GUIs, users did value symmetry in mobile GUIs. In addition, the effect of edge congestion on aesthetics was much weaker. People could seemingly tolerate tighter packed content in mobiles.

Our mediation analysis showed complexity moderated the relationship between aesthetics and clutter, contrast and edge congestion. However, the indirect effect was small and statistically insignificant (Table 9), which generally reflected the overall, relatively weak impact of complexity on aesthetics. It also fitted with our earlier conclusion (Miniukovich & De Angeli, 2014a) that the metrics captured, at least partially, the qualities of GUI relevant to aesthetics, but irrelevant to complexity.

Our best-fit regression models accounted for 40% of complexity score variance and 36% of aesthetics score variance. Including complexity scores in the regression model of aesthetics added only 2% to the explained variance, which could mean we succeeded in accounting for visual-complexity-based aesthetics (Reber, 2012). The performance of our pixel-based metrics was comparable to the performance of metrics from past research. Thus, Purchase et al.'s (2012) metrics explained up to 25% of variation in complexity scores of photographs; Reinecke et al.'s (Reinecke et al., 2013) metrics explained almost 48% of variation in webpage visual appeal scores. However, Reinecke et al. (Reinecke et al., 2013) used several metrics based on idiosyncratic color preferences, which, therefore, could require adjustment for every demographic group. Instead, our metrics were based on the culture-independent effects of perceptual fluency (Reber et al., 2004).

Although we selected test apps from four categories (business, entertainment, travel & local and social apps), we did not conduct a between-category comparison. Our sole purpose was obtaining a sufficient diversity in the apps selected. A reliable between-category comparison would further require sampling significantly more apps than we did in this paper.

2.2.5 Conclusion

This paper presented one of the very first explorations of perceived visual complexity and aesthetics of mobile apps. The results of study 1 suggested people judged mobile app GUIs in the same way as widescreen website GUIs: very quickly and reliably. The results of study 2 suggested people judged mobile interfaces by their visual appearance: six complexity-related metrics explained 40% of their subjective complexity scores and 36% of aesthetics scores. These results suggest the visual quality of mobile app GUIs, as well as of widescreen GUIs (c.f., Miniukovich & De Angeli, 2014a), can be reliably predicted. We plan on embedding the metrics validated in this paper in tLight, our automatic tool for aiding designers (especially non-professional designers) and speeding up GUI development cycles. tLight would quickly test if a GUI was visually simple and appealing, and which dimensions of GUI designs performed poorly. Still, the development of tLight requires further validation and extension of our set of automatic metrics. As the next step, we could develop metrics of grid quality, and conduct metric validation on a larger and more diverse sample of users.

2.3 Computation of Interface Aesthetics

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

2.3.1 INTRODUCTION

There is no doubt visual aesthetics matters in interface design. Surrounded with multiple offers of same-quality services and products, Web and app users have become selective and disregard apps and websites they do not like immediately (Kim & Fesenmaier, 2008). A possible way to survive in such an environment includes carefully working out all details of visual design, making the design stand out (Tractinsky, 2013). However, small companies, start-ups and individual developers often cannot afford hiring a design agency and do their design themselves. In such cases, even well-detailed design guidelines are of limited help, since, to be applied properly, they require extensive training. Concrete GUI evaluation tools could exemplify abstract design guidelines, and drive and substantiate design choices. The tools would be based on specific quality metrics that represent specific GUI design aspects.

In this paper, we extend earlier work (Miniukovich & De Angeli, 2014a; 2014b), and describe and test in two studies eight GUI aesthetics metrics: visual clutter, color range, number of dominant colors, figure-ground contrast, contour congestion, symmetry, and the new metrics of grid quality and white space. We based the metrics on the psychological investigations of what people see as complex and unappealing (Reber et al., 2004), and HCI investigations of webpage aesthetics (Tractinsky et al., 2006; Lindgaard et al., 2006; Tuch et al., 2012; Reinecke et al., 2013; 2014). The results of the present two studies replicated the results of past studies (Miniukovich & De Angeli, 2014a; 2014b), which suggests the metrics are solid enough to be presented to the larger CHI audience. In addition to this, we have replicated the phenomenon of consistent and lasting immediate impressions (Lindgaard et al., 2006; Tractinsky et al., 2006) on two types of stimuli (webpages and mobile apps) using a between-subjects experimental design. In the rest of paper, we review related work on aesthetics in HCI and automatic aesthetics measures, and describe the eight GUI aesthetics metrics. We report Study 1, which tested metric performance on webpages, and Study 2, which tested metric performance on iPhone apps. Lastly, we summarize the results and discuss their implications for the automatic evaluation of interface aesthetics.

2.3.2 Related work

Past attempts (Zheng et al., 2009; Wu et al., 2013; Reinecke et al., 2013; 2014; Miniukovich & De Angeli, 2014b; 2014a) to automatically account for visual aesthetics of GUIs consisted of two steps: gathering user scores of GUI aesthetics and matching them against computed scores of a set of automatic metrics. The first step reflects our understanding that beauty lies in the eye of the beholder, and involves conducting either carefully orchestrated in-lab user studies (Lindgaard et al., 2006; Tractinsky et al., 2006; Purchase et al., 2012; Miniukovich & De Angeli, 2014b) or large-scale crowdsourcing studies with thousands of participants (Reinecke et al., 2013; 2014). The second step uses the averaged user scores as the ground truth data, and tests how well various metrics and algorithms predict the scores.

2.3.2.1 Collecting Aesthetics Scores

The influence of aesthetics on the overall appreciation of GUI changes with time. Sonderegger et al. [29] conducted a longitudinal study and demonstrated the positive effect of aesthetics to almost disappear after the initial use phase. However, it is largely the initial phase that determines if a one-time visitor converts to a user or goes to competitors (Kim & Fesenmaier, 2008). While considering the initial use phase itself, aesthetics impressions could be further subdivided into the *immediate-first* (formed at a glance), *deliberate-first* (long enough for reading titles and processing images) and *overall* (after performing several tasks) impressions (Thielsch et al., 2013).

Several HCI studies have explored the first impression of GUI aesthetics and possible ways of collecting user scores of such impressions. In within-subjects experiments, Lindgaard et al. (2006) and Tractinsky et al. (2006) showed webpage screenshots to participants for half-second intervals, and asked to set a single rating per screenshot. Their experimental design allowed each participant to rate in a relatively short time a large number of stimuli, which, in turn, allowed the authors to make generalizable inferences about stimuli. Tractinsky et al. (2006) then compared the half-second, immediate-first ratings with the ten-second, deliberate-first ratings of each participant and found them to strongly correlate.

Studies (Lee & Koubek, 2010; Thielsch et al., 2013) have compared the deliberate-first and overall aesthetics impressions, and their outcomes. Lee et al. (2010) studied the user preferences of four websites before (deliberate-first impression) and after (overall impression) actual use, and found actual use to significantly change aesthetics appreciation. De Angeli et al. (2006) explored the interplay of various qualities of two website GUIs differing in their level of aesthetics. They asked participants to perform tasks and carefully documented perceived usability, aesthetics and information quality. The effect of aesthetics in GUI appreciation was significant. Notably, both studies (De Angeli et al., 2006; Lee & Koubek, 2010) employed multi-item questionnaires for measuring aesthetics, which, one could argue, is more valid than the one-question approach. However, the whole procedure took several hours for each participant and involved studying only two (De Angeli et al., 2006) and four (Lee & Koubek, 2010) GUIs.

2.3.2.2 Measuring Aesthetics Automatically

Earlier-proposed measures could be generally categorized in *element-based* and *pixel-based*. Element-based measures require knowing the organizational principles (e.g., which GUI elements can contain other elements) and basic elements of GUI (e.g., buttons, links or paragraphs of text). For example, Michailidou et al. (2008) counted the number of menus, images, words and links on webpages, and found these numbers to strongly correlate with the user scores of aesthetics and visual complexity of webpages. Harper et al. (2009) split each webpage in blocks of 200×200 pixels and counted the top-left corners of webpage elements that fell in a block. The distribution of blocks predicted user ranking of 20 webpages with 86% accuracy. Ngo et al. (2003) formulated 14 arts-based measures of graphic display aesthetics. They used the positions and sizes of display elements for quantifying balance, equilibrium, symmetry, sequence, cohesion and nine other webpage characteristics. In a later test of these measures, Purchase et al. (2011) used the HTML sources of 15 webpages to locate text, image and control (e.g., button) elements, and thus, to compute measure scores. A subset of the measures predicted the user ranks of webpage visual appeal. Lastly, Ivory et al. (2001) suggested a number of measures of webpage design quality, which involved counting words, links, images, font types, colors of links and text, text clusters and other webpage aspects. Their metrics predicted if a webpage would be rated very low or very high on visual appeal with 65% accuracy.

Pixel-based measures take GUI screenshots as an input instead of GUI-underlying code. GUI screenshots might represent better what the user sees, which is why pixel-based metrics are considered advantageous to element-based metrics (Reinecke et al., 2013). Zheng et al. (2009) segmented webpage screenshots in color-homogeneous blocks to assess webpage symmetry, balance and equilibrium. A part of their measures correlated with user scores of several aesthetics sub-dimensions. Purchase et al. (2012) explained nearly 25% of variance in user scores of visual complexity of various images. Their measures included computing the variance in image pixel luminance, the ratio of contour pixels to all pixels, file sizes of images in various file formats, and the number of image colors. In two large-scale online studies, Reinecke et al. (2013; 2014) explained up to 49% of variance in webpage aesthetics using the number of text and images blocks, mean values of hue, saturation and value of screenshot pixels, proportion of screenshot pixels of a particular color and several other measures. Finally, Wu et al. (2013) listed a number of measures, both pixel-based

and element-based. The pixel-based measures included decomposing webpages in visual blocks as the first step, and combining the blocks in hierarchical tree-like structures as the second step. Then, the block-average brightness, hue, saturation, texture and colorfulness were computed. These measures, combined with other element-based measures, allowed Wu et al. (2013) to classify webpages with a 77% accuracy, relative to users' classifications. Notably, the webpage decomposition algorithm used in (Wu et al., 2013) was not pixel-based and still required HTML sources of webpages as an input.

2.3.2.3 Complexity Roots of Aesthetics

Psychology (Reber et al., 2004) suggested liking of everyday[5] things might arise from the ease of understanding, i.e., from the simplicity of things. Our brain has evolved to like spending less energy processing things; we see simpler objects as more familiar, and thus, safer. One might argue that complexity is only a single component of liking and conscious considerations might matter even more than complexity (Armstrong & Detweiler-Bedel, 2008). However, the pre-conscious perception of complexity is deemed to be universal (Reber et al., 2004) and is easier to account for than conscious considerations.

Several studies (Choi & Lee, 2012; Nadkarni & Gupta, 2007) have explored what complexity means in HCI and listed three types of complexity: *visual complexity*, *information design complexity* and *task complexity*. Visual complexity strongly relates to immediate aesthetics perception, and many HCI studies focused on exploring and leveraging this type of complexity. Tuch et al. (2012; 2009) have asserted a strong link between visual complexity and immediate aesthetics of webpages. Several other studies (Wu et al., 2013; Reinecke et al., 2013; Reinecke & Gajos, 2014) tried to measure webpage aesthetics automatically, using quantifiable aspects of HTML sources or screenshots of webpages. Despite these studies started from an assumption of a strong complexity-aesthetics link, none of them looked systematically into the underlying determinants of visual complexity, and thus, missed out several potential determinants from their analyses. Earlier work (Miniukovich & De Angeli, 2014a; 2014b) gave a more detailed overview of visual complexity determinants and listed eight of them, relevant to HCI (Table 11).

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

[a] described in (Miniukovich & De Angeli, 2014a; 2014b); [b] introduced in this paper; [c] described in (Miniukovich & De Angeli, 2014a; 2014b), but fully reworked here.

Table 11. The determinants of visual complexity and associated metrics.

2.3.3 Metrics

The metrics of clutter, color range and dominant colors have evolved from the constructs of visual clutter and color variability. Both constructs describe the amount of information on a screen; however, color variability is often measured and considered separately from clutter due to its salience: study participants always mention colors as a component of complexity (Oliva et al., 2004).

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

Visual clutter describes the effort to introduce a new, visually prominent object to a scene (Reinecke et al., 2013) and is quantified with several measures (CL1-CL4, Table 12). Color variability (measures CV1-CV5, Table 12) consists of two aspects that humans perceive separately (Miniukovich & De Angeli, 2014b): number of dominant colors (the colors a human can easily differentiate and name) and color range (the colors a human cannot differentiate without zooming in, and which are often used for smoothing edges and color gradients).

#	Measure description
CL1	Contour density, the ratio of contour pixels to all pixels, cf. (Rosenholtz et al., 2007)
CL2	Subband Entropy (Rosenholtz et al., 2007), the amount of redundancy introduced to a scene.
CL3	Feature Congestion (Rosenholtz et al., 2007), the proportion of unused feature (e.g., color or luminance) variance.
CL4	The file size of images in the JPEG format (cf. (Rosenholtz et al., 2007; Tuch et al., 2009))
CV1	The file size of images in the PNG format.
CV2	The number of colors after color reduction: only values that occurred more than 5 (for web) or 2 (for mobile) times per image were counted.
CV3	The number of colors per dynamic cluster (see CV5).
CV4	The number of static 32-sized color clusters (the sub-cube edge size of clusters is 32 values out of possible 256, per each RGB channel). Only clusters containing more than 5 values are counted.
CV5	The number of dynamic clusters of colors after color reduction (more than 5 pixels). If a difference between two colors in a color cube is less than or equal to 3, two colors are united in the same cluster, which continues recursively for all colors. Only clusters containing more than 5 values are counted.

Table 12. The measures of visual clutter (CL1-4) and color variability (CV1-5).

The metrics of figure-ground contrast and contour congestion reflect the constraints of the human visual system. Figure-ground contrast describes differences in luminance or color of adjacent lines. Smaller differences lead to higher mental effort needed to recognize objects or to read text (Hall & Hanna, 2004). As in (Miniukovich & De Angeli, 2014b; 2014a), we operationalized contrast as a weighted sum of number of contour pixels detected with the Canny algorithm at several consecutive levels. The Canny edge detection algorithm takes two input thresholds: high and low. In the present metric implementation, the low threshold is set to 40% of the high threshold; the high threshold varies from 10% to 70% of maximal pixel luminance (5% to 65% for the mobile app screenshots) with a step of 10%. The weights decrease from 1 to 0 with a step of .2; higher weights are assigned to edge pixels detected with lower thresholds. Finally, the weighted sum is normalized by the number of all detected edge pixels and taken as the metric of contrast.

Contour congestion describes the mental effort needed to differentiate spatially proximal lines. If two objects are too close to each other, a human cannot differentiate them with the peripheral vision and needs to focus on the objects (van den Berg et al., 2009). As in (Miniukovich & De Angeli, 2014b; 2014a), we operationalized contour congestion as the proportion of congested contours to all contours. First, contour pixels are detected. Then, all contour pixels that have neighbors in a 20-pixel vicinity are marked as congested. Finally, the congested pixels are counted and normalized by the number of all edge pixels.

Well-organized information requires less cognitive effort to process. Symmetry (Tuch et al., 2010; Balinsky, 2006) and regular visual layout (Balinsky, 2009) might serve for such a purpose in HCI. In this paper, we reconsidered the past algorithm of GUI symmetry (Miniukovich & De Angeli, 2014b; 2014a) (in which contour symmetry was measured by looking for a match for each contour pixel across the central vertical axis). The past measure was too noisy and favored GUIs with fewer objects. Here, we measured block symmetry, which considers the position of GUI visual blocks,

relative to the central vertical axis. First, we partitioned a GUI screenshot into visual blocks (the algorithm was inspired by (Cao et al., 2010)). Second, we considered separately the blocks that contained the central vertical axis and blocks that did not. The shift of the former relative to the axis was considered as asymmetry. The shift of the latter relative to the axis and matching block (if there was a matching block) was also considered as asymmetry.

The grid quality and white space metrics describe the quality of GUI layout. Higher quality helps the user to quickly navigate within the GUI and is seen as an important aesthetic aspect of GUI (Balinsky, 2009). We considered several existing measures of alignment and regularity of document layouts (Balinsky, 2009), and implemented those that did not require a high precision detector of GUI block positions. We first sliced a GUI screenshots in visual blocks. We then assumed the non-covered proportion of screenshot to reflect badly distributed content and took it as the white space metric. The other five block-based measures (Table 13) describe grid quality and can be combined using a factor analysis. The last measure (Table 13, G5) was added specifically to reflect the specificity of mobile GUIs: they are often organized in a single aligned column of elements, which users are tolerant of scrolling down through. Thus, the measures G1 and G2 should not apply to mobile GUIs.

#	Measure description
G1	The number of visual blocks of GUI (cf. (Balinsky, 2009)).
G2	The number of alignment points of blocks (cf. (Balinsky, 2009)).
G3	The number of block sizes, grid proportionality (Balinsky, 2009).
G4	The proportion of GUI covered by same-size blocks (cf. the cell coverage computation from (Balinsky, 2009)).
G5	The number of vertical block sizes, i.e., vertical grid proportionality.

Table 13. The measures of GUI grid quality.

2.3.4 Study 1

The first study sought to replicate the past results (Miniukovich & De Angeli, 2014a) on a bigger and more diverse sample of websites and pool of participants. In addition, we strived to extend the results from *immediate-first* to *deliberate-first* aesthetics impressions (Thielsch et al., 2013). Past studies of immediate-deliberate differences (van Schaik & Ling, 2009; Tractinsky et al., 2006) used a within-subjects experimental design, whereas we used a between-subjects design: half of participants rated screenshots after a 150ms exposure and the other half after a 4s exposure. After we collected user ratings of aesthetics of 300 webpages, we matched the mean user scores against the scores of the automatic metrics. Additionally, we applied the resulting regression model of aesthetics to stimuli from (Miniukovich & De Angeli, 2014a).

2.3.4.1 Stimuli

A representative sample of stimuli largely determines the validity of findings. Researchers either selected websites themselves (Tuch et al., 2009), relied on online collections of websites preselected for their high quality (Zheng et al., 2009; Miniukovich & De Angeli, 2014a), or asked design professionals to submit website links (Lindgaard et al., 2006). All these approaches implied a sample bias, either because of idiosyncratic preferences of the few people involved in the selection or because of a possible experimenter effect whereby researchers could unconsciously select the "right" stimuli. We, instead, outsourced website search and selection to a larger number of people via crowdsourcing. Over 40 workers of a crowdsourcing platform[6] looked for and reported websites in English of three website genres: corporate, eCommerce and news. We then reviewed over 300 submitted websites and filtered out those that did not match the genres we required, did not have all the page types we required (Table 14), had technical issues or required to login before displaying main content. Keeping in mind that familiarity could seriously complicate data analyses (Tinio & Leder, 2009), we also

[6] https://microworkers.com

avoided the top 500 popular websites[7] and their localized versions (e.g., amazon.it instead of amazon.com). The filtered set contained 235 websites from which we randomly selected 75 websites, 25 in each genre. Finally, we automatically took 300 screenshots of website pages (1280×800 pixels, webpage top part only, PNG format, 24-bit per pixel).

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

Table 14. We took four genre-specific pages of each website.

2.3.4.2 Participants

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

2.3.4.3 Design

We adopted a one-way between-subjects experimental design with exposure duration (150ms vs. 4s) as an independent factor and visual aesthetics as a continuous dependent variable. Participants were randomly assigned to either the 150ms or 4s condition. This manipulation should discern *immediate-first* from *deliberate-first* impressions: 150ms is only long enough to grasp the gist of scene (Fei-Fei et al., 2007) but not enough for reading (Sereno & Rayner, 2003), whereas 4s is long enough for up to 10 eye fixations and reading headlines. The standard duration of 500ms (Tractinsky et al., 2006; Lindgaard et al., 2006) was discarded as it would allow for 1-2 eye fixations and reading 1-2 words, and would mix immediate with deliberate impressions.

2.3.4.4 Procedure

All test sessions were conducted individually, in an isolated room with a laptop and experimenter, and started with a briefing form and consent form. Participants then filled out demographics questionnaires. The study consisted of viewing (1280×800, 13-inch display) and rating 100 webpage screenshots, randomly selected from the pool of 300 screenshots. In each trial (cf., Miniukovich & De Angeli, 2014a), participants saw a fixation cross (1-1.5sec), a webpage screenshot (150ms or 4s, depending on a participant condition), and black-white noise screen (50ms) and were prompted to rate the screenshot with the key buttons from one (ugly) to seven (beautiful). We did not limit the time for rating but explicitly asked participants to do it as quickly as possible. The average completion time was below 15 minutes. At the end of test, participants were debriefed and given a small reward.

2.3.4.5 Results

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

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

The average score interclass correlation coefficients suggested a high consistency in user scores in both 150ms (ICC2k = .77; 95% conf. interval is .74 to 0.81; $F(274, 8970) = 4.95$, $p < .001$) and 4s (ICC2k = .85; 95% conf. interval is 0.83 to 0.88; $F(274, 8671) = 7.08$, $p < .001$) conditions. The analysis of mean scores (from now on we refer to means per screenshot) indicated our sample included both appealing and non-appealing webpages (ranging from 2.1 to 6.1 on a 1 to 7 scale, Figure 6) and was only slightly skewed in both 150ms (n = 275, mean = 3.79, SD = .8, min = 2.1, max = 6.1, skew = .48) and 4s (n = 275, mean = 3.9, SD = .89, min = 1.9, max = 6.3, skew = .24) conditions. A paired t-test revealed a small, but significant difference between the mean scores of 150ms and 4s conditions, diff. = .11, $t(274) = 2.82$, $p < .01$. The correlation between the mean scores of 150ms and 4s conditions was strong, $r(273) = .70$, $p < .001$.

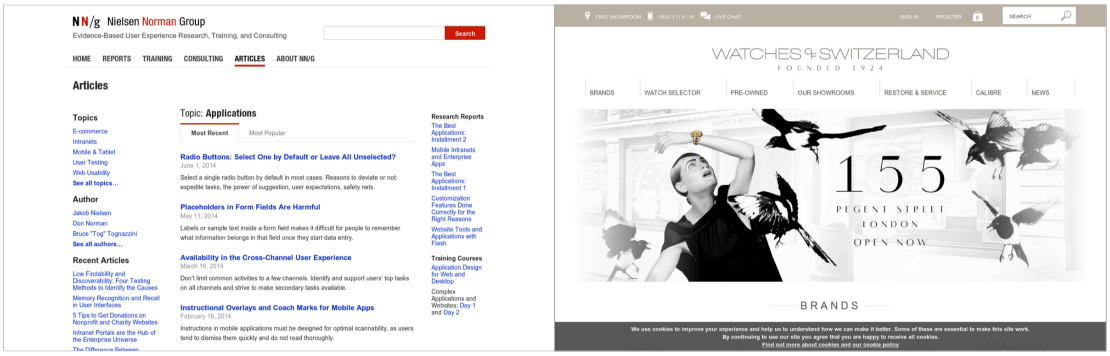


Figure 6. The least appealing (left, 2.1 out of 7) and the most appealing (right, 6.1 out of 7) webpages.

We computed scores of the eight metrics for each screenshot. Four metrics (visual clutter, number of dominant colors, color range and grid quality) required combining multiple measures, for which we used maximum-likelihood factorial analyses. As in (Miniukovich & De Angeli, 2014b; 2014a), we expected the variance of visual clutter and color variability measures to partially overlap and wanted to cancel out the overlapping – we analyzed all the measures in a factor analysis with Varimax rotation (Table 15a). Three factors emerged. Similar to (Miniukovich & De Angeli, 2014b; 2014a), no measures had high cross loadings, with the exception of number of static color clusters (CV4, Table 12), which loaded on both clutter and dominant color factors. We computed Thompson's scores for the three factors. The measures of grid quality (G1-G4, Table 13) were combined in a separate factor analysis (Table 15b), in which a single factor emerged. We again estimated grid quality scores using Thompson's method.

	measures / factors	CL1	CL2	CL3	CL4	CV1	CV2	CV3	CV4	CV5
A	Clutter	.92	.85	.95	.92	.30		-.14	.51	.30
	Color range		.20		.18	.82	.89	.72	.13	
	Dominant colors	.26	.29	.24	.22	.15	.21	-.36	.65	.85
B	measures / factors	G1		G2		G3		G4		
	Grid quality	.77		.67		.90		-.87		

Table 15. Factor loadings of clutter and color measures (A, cumulative var. g = .83) and grid quality measures (B, cumulative var. g = .65) for webpages.

The majority of the automatic metrics generated scores that correlated with the user scores (Table 16). Please note that higher symmetry and grid quality scores correspond to less symmetric and less organized layouts, and therefore, correlate negatively with aesthetics. We then used the Akaike's Information Criterion (AIC) to select the best-fit linear models, which explained 49% (150ms condition) and 43% (4s condition) of user score variance. AIC, in addition to looking at R^2, penalizes models for each additional predictor.

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

Table 16. Pearson's correlation coefficients between metric-produced scores and user scores.

Finally, we applied the factor loadings (Table 15) and linear model coefficients (Table 17, 150ms condition) from this study to the stimuli from (Miniukovich & De Angeli, 2014a), i.e., to a different data set. The amount of explained variance in the aesthetics ratings only dropped down to 42% (from 51% in Miniukovich & De Angeli, 2014a).

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

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

Table 17. Regression models of webpage visual aesthetics. (Outcome variables are aesthetics scores after 150ms and 4s exposure).

2.3.4.6 Discussion

Consistent with the past results (Miniukovich & De Angeli, 2014a), the automatic aesthetics metrics indeed captured certain aspects of webpage visual aesthetics, with clutter and color range being the strongest predictors. Two new metrics (grid quality and white space) also performed fairly well and generated scores that correlated with aesthetics ratings. Our best-fit linear regression model accounted for 49% of variance in the ratings of immediate-first aesthetics, which is comparable to the past results for similar stimuli (Miniukovich & De Angeli, 2014a). Moreover, when applied to a different data set (from Miniukovich & De Angeli, 2014a), the model performed well and explained 42% of rating variance, despite significant differences in data collection of the present and former studies. Notably, we consider the factor loadings (Table 15a) as more valid, compared with the past efforts (Miniukovich & De Angeli, 2014b; 2014a), as they are based on a larger and more diverse sample of screenshots.

Although we cannot directly compare the results (Table 17) with the results of other similar studies (Reinecke & Gajos, 2014; Wu et al., 2013), we would like to point out that the metrics of (Wu et al., 2013), despite an impressive 77% accuracy in predicting user ranking of webpage complexity, performed only 7% better than the Feature Congestion complexity measure (Rosenholtz et al., 2007). Feature Congestion was only one of many measures we integrated in the metrics and when we tried to use it alone to explain aesthetics ratings, R2 dropped to .11 (from .49, Table 17). The webpage aesthetics models of (Reinecke & Gajos, 2014) performed fairly well ($R^2 = .47$), but they included several culture-dependent predictors (e.g., distaste for a particular color) and demographic predictors

(age, gender, geographic location and education level), whereas we used none of them. Collecting demographics and preference data increase GUI evaluation time and cost, and makes an evaluation less affordable to small companies.

Our experimental procedure used a between-subjects experimental design, in which we collected ratings of both *immediate-first* (150ms) and *elaborate-first* (4s) impression of aesthetics. A paired t-test showed elaborate impressions to be slightly more positive than immediate impression (diff. = .11, $p < .01$), which was expected (cf., Reber et al., 1999) and could reflect a lower cognitive strain after watching screenshots for longer time (Tractinsky et al., 2006). The correlation between *immediate-first* and *deliberate-first* scores was strong, $r(298) = .70$, $p < .001$, which supported earlier inferences of within-subjects-design studies (Tractinsky et al., 2006; Lindgaard et al., 2006): participants do form an impression almost instantly and this impression lasts. In fact, several participants of the study explicitly mentioned that the presentation time (4s) could have been shorter and they would still rate screenshots the same. When we used the scores of immediate and elaborate impressions in regression analyses, the R^2 of the resulting models (Table 17) only differed by 6%. This seems to suggest that people rely on similar GUI aspects in judging beauty, regardless of exposure duration.

2.3.5 Study 2

A previous study (Miniukovich & De Angeli, 2014b) argued that pixel-based complexity metrics could successfully evaluate any type of GUIs, not only webpages. Study 2 sought to support the argument and replicate past studies (Miniukovich & De Angeli, 2014b, study 1 and 2) on a bigger and more diverse sample of stimuli (iPhone apps) and pool of participants. The design of Study 2 resembled the design of Study 1; hence we only highlight the differences between the two.

2.3.5.1 Stimuli

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

Instead, we crawled the Apple's app store website and selected free apps in English from three genres: business, travel and entertainment. These genres should represent well the whole of the task-fun continuum, with business apps being mainly task-oriented, entertainment apps being mainly fun-oriented, and travel apps occupying the middle. After we selected three lists of apps, we randomized the app order and considered the apps one by one from the top according to several criteria. First, we only considered apps designed for the 4-inch displays (the apps for smaller displays would render with two black areas at the top and bottom of display). Second, we avoided overly simplistic apps that had less than four visually diverse GUI layouts. Third, we avoided apps with the heavy use of cartoon graphics. Forth, we avoided apps that required a paid premium account to access their full functionality. Last, to reduce possible familiarity effects, we avoided the apps from the top 100 most popular apps list. After randomly selecting 25 apps per genre, we manually took four or more screenshots (640×1136 pixels; PNG format, 24-bit per pixel) per app and then randomly reduced to only four screenshots per app. Thus, we collected 300 screenshots of 75 iPhone apps.

2.3.5.2 Participants

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

2.3.5.3 Design & Procedure

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

2.3.5.4 Results

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

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

As in Study 1, we computed scores for the eight metrics, four of which (visual clutter, number of dominant colors, color range and grid quality) needed additional maximum-likelihood factorial analyses to combine multiple measures. Three factors emerged in the factor analysis with Varimax rotation of clutter and color measures (Table 18a). One factor emerged (Table 18b) in the factor analysis of the grid quality measures relevant to mobile GUIs (G3-G5, Table 13). We then estimated the scores of corresponding metrics using Thompson's method.

	measures / factors	CL1	CL2	CL3	CL4	CV1	CV2	CV3	CV4	CV5
A	Clutter	.94	.86	.94	.89	.19	.14		.22	
	Color range	.10	.17	-.11	.31	.91	.76	.70	.22	
	Dominant colors		.23		.15	.35	.43		.69	.99
B	measures / factors	G3			G4			G5		
	Grid quality	.99			-.64			.86		

Table 18. Factor loadings of clutter and color measures (A, cumulative var. g = .82) and grid quality measures (B, cumulative var. g = .72) for iPhone apps.

The output of the symmetry metric was not normally distributed, since a large part of app layouts were almost perfectly symmetrical. We converted the output of the symmetry metric in a categorical variable: screenshots with a score < 10 were considered symmetrical (104 app layouts); screenshots with a score > 200 were considered asymmetrical (90 app layouts); the rest was considered as partially symmetrical (106 app layouts). Screenshots of the same app often varied widely on symmetry (Figure 7). A one-way ANOVA test revealed a significant effect of symmetry on user score in 4s condition ($F(2,297) = 4.62$, $p < .05$), but not in the 150ms condition ($F(2,297) = 1.01$, $p = .37$). Post-hoc comparisons using the Tukey HSD test showed participants disliked partially symmetrical layouts relative to fully symmetrical and asymmetrical layouts (Table 19b). For the rest of metrics, we estimated the correlations between their output and user scores (Table 19a). The best-fit linear models (based on the Akaike's Information Criterion) explained 13% (150ms condition) and 18% (4s condition) of user score variance (Table 20). When the effect of app genre was considered, the fit of the models went up (17% in the 150ms condition and 32% in the 4s condition).

We also applied the factor loadings (Table 18) and linear model coefficients (Table 20, 150ms condition) from this study to the Android app screenshots from (Miniukovich & De Angeli, 2014b). The amount of explained variance in the Android app aesthetics only dropped down to 30% relative to 36% in (Miniukovich & De Angeli, 2014b).

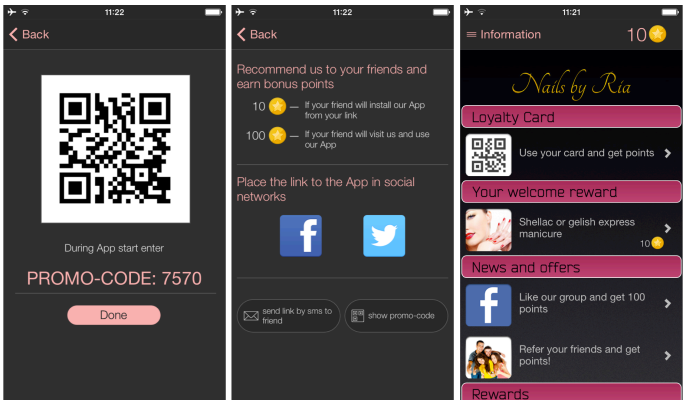


Figure 7. Symmetrical (left), partially symmetrical (center) and asymmetrical (right) screen layouts of an app.

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

Table 19. Pearson's correlation coefficients of user scores and continuous metrics, and between-level differences for the categorical metric of symmetry.

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

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

Table 20. Linear regression models of app visual aesthetics. (Outcomes are aesthetics scores after 150ms or 4s exposure).

2.3.5.5 Discussion

Overall, the results of study 2 confirmed our expectations: the majority of the automatic metrics generated scores correlating with the user ratings of iPhone app aesthetics (Table 19). Compared with a similar study on Android apps (Miniukovich & De Angeli, 2014b), the best-fit models (Table

20) were less effective at explaining iPhone app aesthetics. However, when applied to the past dataset (Miniukovich & De Angeli, 2014b), the present study model (Table 20, the 150ms condition) explained 30% of Android app aesthetics, which demonstrated the reliability of the metrics and suggested they can be applied to both web and mobile GUIs. We consider the factor loadings (Table 18) as more valid compared with our past efforts (Miniukovich & De Angeli, 2014b), as they are based on a larger and more diverse sample of screenshots. The inclusion of app genre in the models increases R^2 (up to .32 in the 4s condition), but significantly complicates the analysis. Further studies are needed.

Study 2 used between-subjects experimental design, comparing the ratings of both *immediate-first* (150ms) and *deliberate-first* (4s) impressions of app beauty. As in Study 1, a paired t-test showed elaborate impressions to be slightly more positive than immediate impressions (diff. = .17, $p < .001$), which was also consistent with the past research on affective judgments (Reber et al., 1999). However, the correlation between immediate and deliberate impression ratings remained strong ($r = .58$, $p < .001$). This further supports the claims of aesthetics impressions forming immediately and lasting (Tractinsky et al., 2006; Lindgaard et al., 2006) and extends them from website-only to mobile GUIs as well.

2.3.6 General Discussion

Study 1 and Study 2 largely converge. Using between-subjects experimental design, both studies observed a strong link between *immediate-first* and *deliberate-first* aesthetics impressions (Table 21a). This reinforces past findings (van Schaik & Ling, 2009; Tractinsky et al., 2006) and extends them from the web to the mobile domain (see also Miniukovich & De Angeli, 2014b, study 1).

Passing from immediate-first to deliberate-first impression brings us closer to the real-world usage situations. (One might even argue immediate impressions are only possible in a lab.) The automatic metrics were initially designed to account only for low-level, perceptual GUI qualities, and not for high-level conscious elaborations. However, reading titles and considering images do not drastically change user scores (Table 21a) and performance of the metrics (Tables 6 & 9). We assume that the influence of GUI visual aspects carries over from the very initial, 150ms-long phase to a more elaborate, 4s-long phase.

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

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

Table 21. Comparison of results of studies on web and mobile GUIs.

Predicting GUI aesthetics has proven to be harder for mobile apps than for webpages, which the comparison of R^2 of regression models demonstrates clearly (Table 21c). We could attribute this to two reasons. First, visual complexity might matter much less to mobile app than to webpage evaluators, either because all apps are already designed with complexity concerns in mind (e.g., for the on-the-go use or difficult lighting conditions) or because smaller screens imply fewer details to perceive. Thus, the complexity-rooted metrics explain less of mobile aesthetics. Second, Apple reviews all iPhone apps published on their app store, which means the apps are already preselected

based, in part, on their design. The iOS design guidelines[8] advices developers to *"use a family of pure, clean system colors that look good at every tint …", "align text, images, and buttons to show users how information is related"* or to *"create a layout that fits the screen of an iOS device."* Following the guidelines restricts the variance in the scores of several metrics, which also reduces the chances for covariance, i.e. for stronger correlations. Indeed, the metric performance triples on Android apps (R^2 = .30), which Google Play Store publishes without review. We call for more studies on mobile apps, as they are becoming increasingly more important, and yet, rarely mentioned in literature.

The web-mobile comparison of metric performance (Table 21b) also revealed significant resemblance. Except white space, symmetry and dominant color, all metrics functioned similarly (in both, correlation direction and magnitude) for both web and mobile GUIs. The effect of dominant colors did not reach a significant level, which could reflect the overall tendency to use fewer dominant colors in mobile apps: restricted variance in dominant colors would lead to limited chances for covariance with aesthetics. The white space metric did not apply to mobile GUIs, which could follow from participants tolerating well incomplete list- or menu-like GUIs of many mobile apps. Finally, the output of symmetry metric for iPhone apps was not normally distributed (one-column layouts of apps often imply full symmetry) and was converted to a categorical variable with three levels (full symmetry, partial symmetry, and complete asymmetry). Instead of a linear drop in liking from full symmetry to complete asymmetry, we observed partial symmetry to be disliked the most (Table 19b), which was in part consistent with the recent research on slight asymmetries (Gartus & Leder, 2013). This might also mean complete asymmetry was perceived as an intended feature and tolerated. Overall, we conclude that the same factors matter for both web and mobile GUI aesthetics, and accordingly, the metric can be similarly applied to both web and mobile GUIs.

2.3.7 Conclusion

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

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

3 MODEL APPLICATION

3.1 Computational Aesthetics: From Webpages to Websites

Computational aesthetics aims at modeling and predicting visual aesthetics *automatically*, which could help HCI researchers understand better user preferences and HCI practitioners create better websites. Computational-aesthetics research has developed the methods to evaluate webpage aesthetics and demonstrated the method validity and reliability. However, webpages differ from websites. Webpages combine to form a website, are interacted with in combination, and may have to be studied in combination, not as separate entities. This paper investigates the computation of website, holistic aesthetics. A user study validated our adaptation of for-webpage computational method to websites and explored two additional, website-level factors of aesthetics, visual diversity and conceptual complexity. The results of study suggested that the adapted computational method could indeed predict the user scores of aesthetics. Visual diversity and conceptual complexity did appear to describe unique aspects of website aesthetics, however, we call on further research to validate such result and incorporate these two aspects in aesthetics computation. The paper concludes by discussing suggestions for the future research on computational website aesthetics.

3.1.1 Introduction

Aesthetics adds value to information systems. Aesthetically pleasing systems have higher hedonic appeal, which attracts and excites users, and makes interaction fun. From the business perspective, higher aesthetics results in higher customer satisfaction and intention to repurchase (Kim et al., 2011). Even more crucial, aesthetics often serves as a differentiating factor between successful and failing systems (Tractinsky, 2013); a non-appealing system fails to grasp user attention, who abandons the system to the benefit of its rivals. For example, a typical flow of action in the website domain may go as follows (Kim & Fesenmaier, 2008). Users first arrive to a search-engine website. They formulate and send a query to the engine, and receive a list of results, short text snippets describing potential query-matching websites. The users select a link, go on the site and have a brief look – more of a peek – at the website. If the impression is not good enough, the user does not invest more time in browsing and opens a rival website from the search-result list. Aesthetics principally determines first impression (Thielsch et al., 2013; Sonderegger et al., 2012; Lindgaard et al., 2011); aesthetics keeps new users coming to the site.

In the last decades, HCI research on aesthetics has flourished. A vast number of user studies have been run asking people to fill in standardized questionnaires (Tractinsky & Zmiri, 2006; Moshagen et al., 2009) or to engage in interviews (Moshagen & Thielsch, 2010). This corpus of research has led to the definition of some design guidelines (Sutcliffe, 2002; Sutcliffe, 2009) and more recently to a promising niche of research on automating aesthetics evaluation (Reinecke & Gajos, 2014; Reinecke et al., 2013; Wu et al., 2011; Wu et al., 2013; Zheng et al., 2009; Altaboli & Lin, 2011; Purchase et al., 2011; Miniukovich & De Angeli, 2014a; 2014b; 2015a). Aesthetics computation may not be able to replace user studies (cf., Rosenholtz et al., 2011; Purchase et al., 2012), but it can quickly give numerical and visual feedback, which fits well quick iterative development imposed by current market (cf., Altaboli & Lin, 2011, p.2).

Most computational studies used static webpages as a unit of analysis to measure aesthetics. However, static pages might not approximate websites, e.g., they do not convey on-page dynamics (e.g., unfolding menus) and cross-page dynamics (visual difference between pages) of websites. The presence of such dynamics reflects a specificity of aesthetics research on graphical user interfaces (GUIs) – it explores stimuli that morph and change. On the contrary, other research fields such as document aesthetics (Harrington et al., 2004; Balinsky, 2009) or image aesthetics (Datta et al., 2006) explore inherently-static documents and images. The dynamics and other specificities of GUIs have been discussed in HCI, starting from the differentiation between classical aesthetics and expressive aesthetics, which was proposed and later criticized by Tractinsky (Porat & Tractinsky, 2012; Tractinsky, 2013; Lavie & Tractinsky, 2004). However, the GUI specificities have been rarely accounted for computationally, with the majority of research treating GUIs as disconnected collections of static layouts.

We aspire to advance the work on aesthetics computation and to move on from screenshot aesthetics to website aesthetics. In previous work (Miniukovich & De Angeli, 2014a; 2014b; 2015b), we proposed a method of static-GUI aesthetics computation. Here, we summarize the method and adapt it to be applicable to websites. The adaptation included solving such issues as combining the aesthetics of individual pages in a holistic score, and the effect of website conceptual complexity and cross-page visual diversity on aesthetics. The method was validated in a user study, in which 45 participants browsed, ranked and rated 30 websites before and after use. The computed scores correlated with the reported pre-use and post-use aesthetics, which meant our algorithms could differentiate beautiful from ugly websites. We consider this result an interesting development of the aesthetics computation methods, and invite their further discussion and exploration.

In this paper, we review related work in section 2, describe the past method for webpage aesthetics computation and the new method for websites in section 3, describe a user study to validate the adapted method in section 4, summarize study results in section 5, discuss the results in section 6, and conclude by outlining future challenges to GUI aesthetics computation in section 7.

3.1.2 Related Work

A large number of people interact with most GUIs in a discretionary-use context; they are not forced to use a GUI and can switch to alternative GUIs at will. In such discretionary use, aesthetics largely determines if a GUI gets chosen from the multitude of alternatives (Diefenbach & Hassenzahl, 2011). The impact of aesthetics on user choices occurs at the beginning of interaction, when the user forms a quick impression about system quality (Karapanos et al., 2009; Sonderegger et al., 2012)

Upon the initial encounter with a GUI, user impression evolves through three stages: *immediate*, *deliberate* and *post-use* impression (cf., (van Schaik & Ling, 2009; Thielsch et al., 2013)). Stable immediate impression forms within the first 500ms (Lindgaard et al., 2006; Lindgaard et al., 2011; Tractinsky et al., 2006) and carry over into deliberate impressions (several seconds of title reading and image watching, (Tractinsky et al., 2006; van Schaik & Ling, 2009), which, in turn, carry over into the final, post-use impressions. With each consecutive step, more cognitive reasoning affects the overall impression and user choices, diluting the effect of immediate impression (van Schaik & Ling, 2009). However, the effect of aesthetics on user judgment never completely vanishes and stays strong even after participants complete several tasks (Lee & Koubek, 2010; Sonderegger & Sauer, 2010; Thielsch et al., 2013).

Most computational-aesthetics studies in HCI focused on immediate and deliberate impressions (Reinecke & Gajos, 2014; Reinecke et al., 2013; Wu et al., 2011; 2013; Zheng et al., 2009; Altaboli & Lin, 2011; Purchase et al., 2011; Miniukovich & De Angeli, 2014a; 2014b; 2015a). The studies needed to collect enough data for correlational analyses (Hassenzahl & Monk, 2010), and thus, required participants to rate large numbers of stimuli. However, participants could only rate many stimuli after passive viewing, not after active browsing, which meant the studies collected viewing-based immediate and deliberate impressions. Collecting the post-use impressions for many stimuli constitutes a challenge. As a consequence, the majority of active-browsing studies investigated few websites, two to four (Hartmann et al., 2007; Thielsch et al., 2013; De Angeli et al., 2006). Nonetheless, post-use impression corresponds closely to the real-world impression and computational aesthetics should strive to develop and validate the methods capable of estimating the post-use impression.

3.1.2.1 Aesthetics Computation

HCI researchers tried out multiple approaches to aesthetics computation. The majority of research can be categorized along two dimensions: theory and operationalization (Figure 8). Theory describes the epistemology behind the algorithms and varies on a continuum between the art and psychology (c.f., Shimamura, 2012; Silvia, 2012). On the one hand, art historians and philosophers have long been debating the rules of aesthetics (e.g., the rule of thirds or golden ratio). Such rules may well reflect human aesthetic preferences. On the other hand, psychologists have observed exploitable patterns in human's appreciation of things (e.g., preference for symmetry). Both arts and psychology have inspired a number of works in HCI.

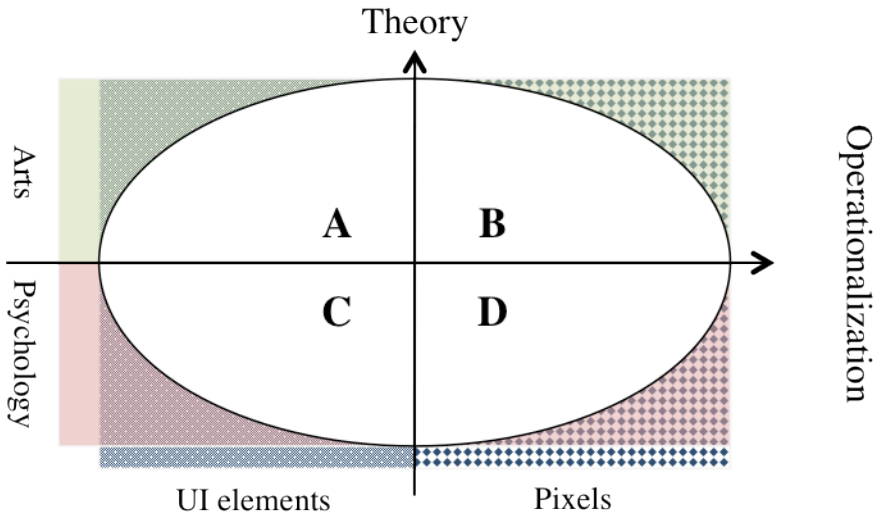


Figure 8. The foundations of different approaches to computational aesthetics in HCI.

The operationalization dimension describes how the algorithms measure aesthetics and also varies between two poles. The first pole implies basing algorithms on the standard elements of GUIs (e.g., buttons, icons or text labels) and their visual properties (e.g., color, size or spatial position). Such element-based algorithms often take GUI underlying codes (e.g., HTML and CSS in the case of websites) as an input and treat all GUI elements as rectangular "blocks", for simplicity. The other pole implies basing algorithms on the pixels of GUI screenshots, which makes them applicable without substantial adaptation to any GUI kind (e.g., web, mobile or ATM terminal). Processing screenshots instead of GUI elements may increase the complexity of algorithms, but it may also improve their performance, since screenshots represent exactly what a human sees, whereas GUI codes do not (cf., Reinecke et al., 2013). In addition, designers often create GUI concepts, such as images or wireframes, without generating valid, optimized GUI markup. At such an early design stage, only pixel-based methods can provide feedback on the concepts, since they use image processing.

The research in quadrant A (Figure 8) was brought about by Ngo et al. (2003; Ngo, 2001), who proposed and initially tested 14 art-based measures of layout aesthetics. Out of these 14 measures, Altaboli & Lin (2011) computed three (balance, unity and sequence) for abstract screen models and websites. The measures of layout unity and sequence predicted well user scores of webpage classical aesthetics (Lavie & Tractinsky, 2004) and simplicity (one of aesthetics dimension from Moshagen & Thielsch, 2010). Purchase et al. (2011) computed all 14 of Ngo et al.'s (2003) measures for 15 webpages, based on the positions and sizes of texts, images and control elements (e.g., buttons). Several measures predicted the user ranks of webpage appeal. Finally, the work of Helen Balinsky (2006) also belonged to quadrant A. She described a sophisticated measure of symmetry in documents. The measure operated on the rectangular blocks of content (not pixels) and was presented as a part of document aesthetics.

The intersection of art theories and image processing represents quadrant B, Figure 8. Zheng et al. (2009) used the quadtree image decomposition to implement three out of 14 Ngo et al.'s (2003) metrics, namely symmetry, balance and equilibrium. The computed scores of symmetry and balance corresponded to user evaluations of 27 webpages. Bhattacharya et al. (2010) suggested algorithms for photograph enhancement, which were based on the rule of third and golden ratio. In more detail, they suggested to position the focus of visual attention (foreground objects on the photographs) at the distance of a third of photograph width (height) from the photograph edges, and to rescale the ratio of empty areas (e.g., skies in a photograph) to content areas (e.g., houses or lands) according to the golden ratio. Bhattacharya et al. (2010) reported 73% of enhanced photographs to be preferred over the original photographs.

Many HCI researchers exploited psychology knowledge – complex things require a high processing effort and are generally disliked (Reber et al., 2004; Reber, 2012). Some of these

researchers processed GUI elements: their work belongs to quadrant C, Figure 8. Michailidou et al. (2008) counted webpage menus, images, words, links, and top-left corners of all visible elements. The counts correlated with the user scores of visual complexity, but not with the scores of visual aesthetics. Harper et al. (2013) further explored the counts of top-left corners of webpage visible elements: they counted corners falling in 200 x 200 pixel squares. The distribution of corners across the squares correlated strongly with user ranks of 20 webpages. For a large sample of 1898 webpages, Ivory et al. (2001) calculated a number of measures: size of pages in bytes; ratio of title and emphasized text to all text; number of links, word, text clusters, fonts, images, colors of texts and links; and several other measures. Depending on a website genre, these measures predicted from 11% to 56% of website quality ratings coming from the judges of the 2000 Webby Awards competition. Wu et al. (2011; 2013) applied machine learning to classify webpages, using both element-based (layout and text features, e.g., the number and sizes of layout blocks, the number of texts, or character density) and pixel-based features (color, texture and visual complexity features, e.g., pixel hue, brightness, saturation and colorfulness, and screenshot file sizes in the JPEG format). Classification accuracy reached 77%. Such result might seem superior to other results, but we have to remember that classifying differs from fitting a linear model as done in other papers.

Several other researchers also relied on visual complexity as the premise of their algorithms, but borrowed their methods from image processing. They represent quadrant D, Figure 8. Purchase et al. (2012) explained 25% of perceived visual complexity of photographs with the number of (non-)reduced colors, variation in photograph pixel luminance, ratio of contour pixels to all pixels, and sizes photograph files in the JPEG, PNG and GIF formats. Reinecke et al. (2014; 2013) predicted up to 49% of webpage visual appeal. In addition to several demographic variables (e.g., age, country of residence, or education level), their list of predictors included the average pixel hue, saturation and value, colorfulness (Hasler & Suesstrunk, 2003), the number of images and text areas, layout symmetry, balance and equilibrium, and several other measures. Miniukovich & De Angeli (2014a; 2014b; 2015a) proposed eight pixel-based measures, which accounted for up to 51% of webpage visual aesthetics and 36% of mobile app visual aesthetics. This paper reports a study that falls in quadrant D.

3.1.2.2 Webpages VS Websites

Unlike the computation of webpage aesthetics, the computation of website aesthetics has not been properly explored. Researchers often described results using the terms *website* and *webpage* interchangeably (e.g., in (Kim & Fesenmaier, 2008), despite they studied static webpages or webpage screenshots. In reality, users interact and form opinions about websites, not discrete webpages.

Website aesthetics differs from webpage aesthetics in several regards, and an analysis of website aesthetics should consider website-level factors (cf., (Ivory & Hearst, 2002). First, a score of website aesthetics may need to combine the scores of the multiple webpages that constitute the website; users may well go beyond a homepage and be affected by other pages (cf. (van Schaik & Ling, 2009)). Little information exists on how to compute this, which pages should be considered and if the pages should have the same or different weights in an aggregation formula are still open questions. Many researchers solely considered homepages, since the user sees them first and is impacted by them the most (e.g., Kim & Fesenmaier, 2008; van Schaik & Ling, 2009), but such strategy is yet to be justified.

Second, website aesthetics may depend on cross-page visual diversity, the opposite of visual consistency. Archambault & Purchase (2013) studied mental map preservation in dynamic graphs – a concept related to being visually consistent. The concept of mental map preservation is widely used in visualization literature; it refers to changing as few aspects as possible during animation, and thus, reducing mental load. The study (Archambault & Purchase, 2013) suggested that preserving one's mental map improved the user performance on within-graph orientation tasks. Moore et al. (2005) explored the consistency between a webpage and banner ad in user studies. Their results showed that people better noticed and recalled page-inconsistent banners, but preferred consistent or slightly inconsistent banners. Ivory & Hearst (2002) automatically estimated the diversity of pages across a site. They computed the scores of page appearance quality and took the standard deviation of scores normalized by the score mean as a diversity metric. Computing such metric required processing at least five pages per website.

Van der Geest & Loorbach (2005) explored visual diversity in web sites in a user study. Their participants sorted webpages in groups and labeled the groups. The study highlighted the high

importance of *visual aspects* in user perception of website diversity: the participants used visual-aspect descriptions 59% of the time – much more than associative (7%), function-related (19%) or other descriptions (16%). The descriptive labels related to six visual aspects: color, background, font, illustration, grid/navigation and logo. The color and grid/navigation visual aspects accounted for 166 out of 210 labels. Van der Geest & Loorbach (2005) advocated keeping webpage look strictly consistent within sites, though admitted slight diversity might have served designer's goal "*to indicate that the user is entering a new content area in a site*". Miniukovich & De Angeli (2015b) automatically estimated the visual diversity of webpages and mobile apps. The estimation required processing at least four webpages or app layout, and showed visual diversity to correlate with the metrics of success, such the number of users installing or commenting on an app.

Finally, the aesthetics impression of a website may depend on the quality of interaction, and by extension, on the website structure: as the user moves from one webpage to another, the experience of surfing affects the user more and more, with the effort of sensemaking of website structure being the key contributor. Nadkarni & Gupta (2007) and Choi & Lee (2012) addressed the sensemaking effort via the concept of coordinative (aka, conceptual) complexity, which described the complexity of information clusters of websites and mobile apps. If we considered webpages as information clusters, then website-level complexity could be represented by the number of pages and links, average webpage length, and various descriptors of structure of website page graph (cf., graph linearity and optimal navigation path, (Gwizdka & Spence, 2006). Such website-level aspects may well be included in the automatic evaluation of aesthetics impression.

3.1.3 Website Aesthetics Computation

We based our research on psychological theories and pixel-based operationalization of GUI aesthetics (Figure 8, quadrant D). The alternative, arts-based approaches to aesthetics (Ngo, 2001; Ngo et al., 2003) have received much attention in HCI, but they have lacked theoretical backing and their effectiveness was not convincingly demonstrated. We noticed that stimuli were highly artificial (e.g., black squares on a white screen, (Altaboli & Lin, 2011) or too few (e.g., five GUI screens in Ngo et al., 2003) and 15 webpages in (Purchase et al., 2011), or only a small subset of metrics was tested (e.g., three in Zheng et al., 2009).

The present work approaches the computation of website aesthetics as a two-part problem: assessment of individual webpages, and assessment of websites as holistic artefacts. Part 1 leverages the screenshot-based method of aesthetics computation for individual webpages (Miniukovich & De Angeli, 2014a; 2014b; 2015a). Part 2 builds on the original contribution of this paper: adaptation of the screenshot-based method for websites and consideration of new, website-level factors of aesthetics. We then validated this two-part approach with a user study.

3.1.3.1 Webpage Evaluation

The aesthetics measures from (Miniukovich & De Angeli, 2014a; 2015a) consisted of color range, dominant colors, contour congestion, figure-ground contrast, grid quality, white space, symmetry, and visual clutter. Past research (Rosenholtz et al., 2007; Purchase et al., 2012; Balinsky, 2009) suggested multiple measures for visual clutter, color and grid quality. We combined several of them in composite metrics for increased reliability. The other measures – contour congestion, contrast, white space, and symmetry – are single-item. We refer the reader to Miniukovich & De Angeli (2014a; 2014b; 2015a) for a deeper description of the method and present only a brief summary of it in this paper.

Visual clutter describes the number, density and dissimilarity of objects in a visual scene. Study participants mention most often the high number of objects when describing complex scenes (Oliva et al., 2004), and psychologists often called the number of objects as a set size (Rosenholtz et al., 2007) and viewed it as the main component of complexity (Tinio & Leder, 2009; Bravo & Farid, 2004). Our metric of clutter combined four sub-measures: *contour distribution* (Figure 9), which measures the ratio of contour pixels to all pixels and roughly corresponds to the number of objects on a screen; *Subband Entropy* (Rosenholtz et al., 2007), which describes the amount of non-essential, often repetitive, information in a scene; *Feature Congestion*, which describes the visual uniqueness of scene objects; and *file sizes of JPEG-compressed images*.

Dominant colors (Figure 10) represent the image colors that a user can easily see and count. A higher number of colors implies higher dissimilarity in a visual scene (Bravo & Farid, 2004), and by

extension, higher cognitive load. Participants invariably named colors as a complexity component (Oliva et al., 2004), which made dominant colors worth differentiating from visual clutter in a separate metric. The sub-measures of dominant colors included the number of static color clusters and number of dynamic color clusters (cf., Figure 10).

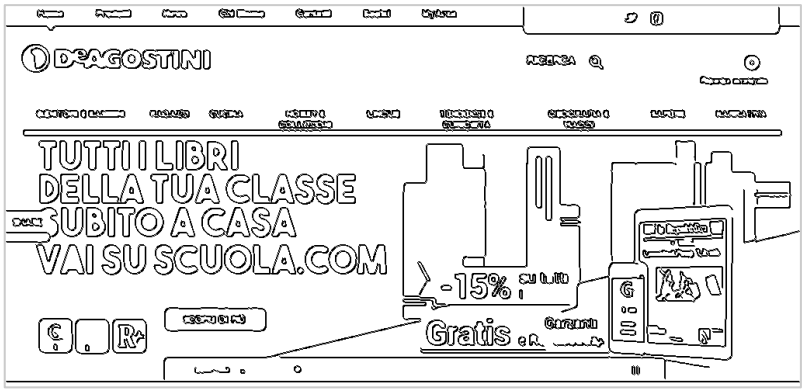


Figure 9. Webpage contours. More contour pixels (relative to a screenshot size) correspond to higher visual complexity, and thus, lower aesthetics.

Figure 10. Dominant colors. The number of dominant colors and proportion of page that the colors occupy are shown under the two example pages.

Color range describes the whole range of webpage colors, including the cohort of shades and semitones that come along dominant colors (Figure 11). Human vision discards gradual changes in color (gradients and shadows) as belonging to the same color and focuses on sharp changes (often corresponding to contours), (Rizzi & McCann, 2007). However, gradients and shadows naturally occur around us; the resemblance to known, naturally occurring scenes increases liking and aesthetic value. The metric of color range combined three sub-measures: the number of colors (RGB 24-bit values) after color reduction; the average number of colors in dynamic color cluster; the file sizes of images in the PNG format (PNG file sizes grow with each additional color).

Contour congestion quantifies the density of image contours (Figure 12). To be recognized, overlapping or adjacent small objects require participants to focus on them. Having more of such objects increases perceived scene complexity (van den Berg et al., 2009). The measure of congestion estimated the ratio of congested contours to all contours. Contours are first detected using a Sobel-like edge detection algorithm, both horizontal and vertical. Contour pixels are then marked as congested if another contour appears in their 20-pixel vicinity.

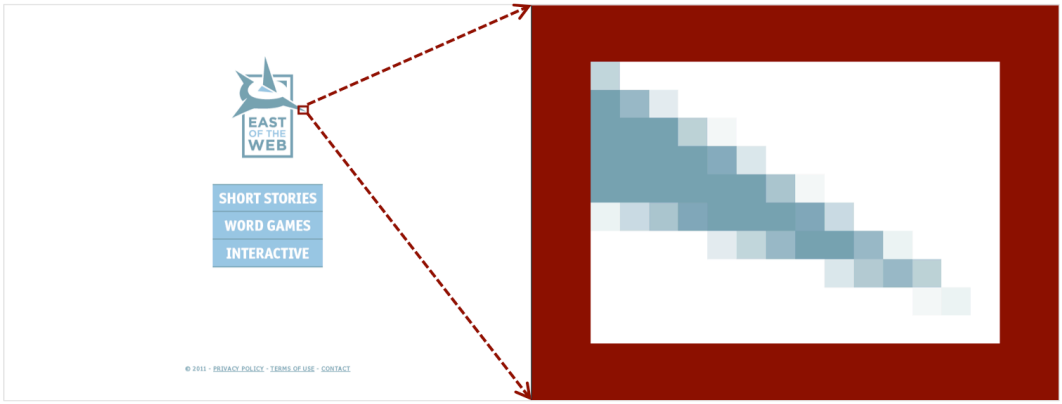


Figure 11. A two-color webpage? Zooming in (on the right) shows the multitude of color shades that come along with each dominant color.

Figure 12. Contour congestion. Too dense, congested contours are red; the other contours are green.

Figure-ground contrast (Figure 13) describes a difference in luminance or color between adjacent objects. Psychologists have long been using higher contrast to facilitate reading (Hall & Hanna, 2004) and object recognition (Reber et al., 1999). Accordingly, the measure of contrast assigns higher weights to fainter contours. We detected image contours at 6 levels of rigidness. We then counted the number of contour pixels at each level, multiplied the counts by the level weights, and normalized the final sum by the number of all edge pixels.

Grid quality corresponds to the regularity of GUI grid. Designers often fit webpage content in a grid: they partition the content in rectangular blocks and align the blocks to a few vertical alignment points (such a point corresponds to a point where a GUI block starts or ends.). Higher grid regularity (or repetitiveness) simplifies a GUI: research on website design suggested using fewer blocks, using blocks of the same size, and aligning the blocks to the same alignment points (Balinsky, 2009; Harrington et al., 2004). The grid quality metric first segments a GUI screenshot in rectangular blocks (image edges are detected and fitted in rectangular blocks; edge-free areas or lengthy straight lines separate different blocks; tiny blocks are discarded). The segmentation then serves a basis for four sub-measures of grid quality: the number of GUI blocks; the number of alignment points (Figure 14); the number of block sizes; and the proportion of GUI taken by the blocks of prevalent size.

Figure 13. Contrast measure. The yellower contours are rigid; the bluer contours are faint. (A part of original webpage is on the right).

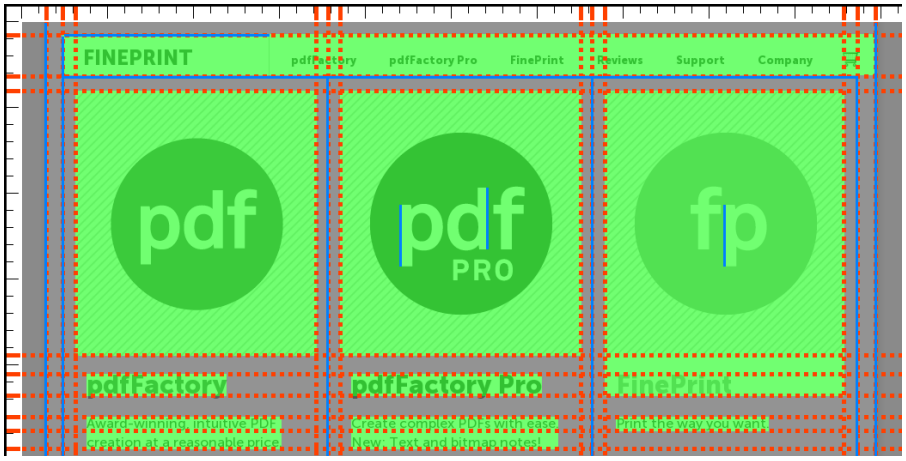


Figure 14. Alignment points. Webpage content is segmented in blocks (green); blue lines show impactful straight lines; dotted red lines show the alignment points. (The background webpage was decolorized.)

White space describes the proportion of GUI that is free of text, graphics and functional elements. HCI research has observed that excessive white space might correspond to GUI layout issues (Riva et al., 2010; Harrington et al., 2004; Miniukovich & De Angeli, 2015a) and decrease GUI aesthetics. The measure of white space segments GUIs in rectangular blocks (the same algorithm as in Grid Quality) and calculates the ratio of between-block space to the whole GUI size.

Symmetry – vertical mirror symmetry (Figure 15) – facilitates object perception (Machilsen et al., 2009), and thus, makes the objects more beautiful. Notably, HCI research provided mixed evidence of such principle applying to GUIs, probably reflecting the difficulty of measuring the symmetry of highly-complex realistic GUIs, (Bauerly & Liu, 2006; Tuch et al., 2010). The symmetry metric segments a GUI in rectangular blocks and estimates the size of blocks – or parts of blocks – unmatched across the central vertical axis. The sum of such unmatched, asymmetrical blocks is normalized by the sum of all blocks and taken as the measure.

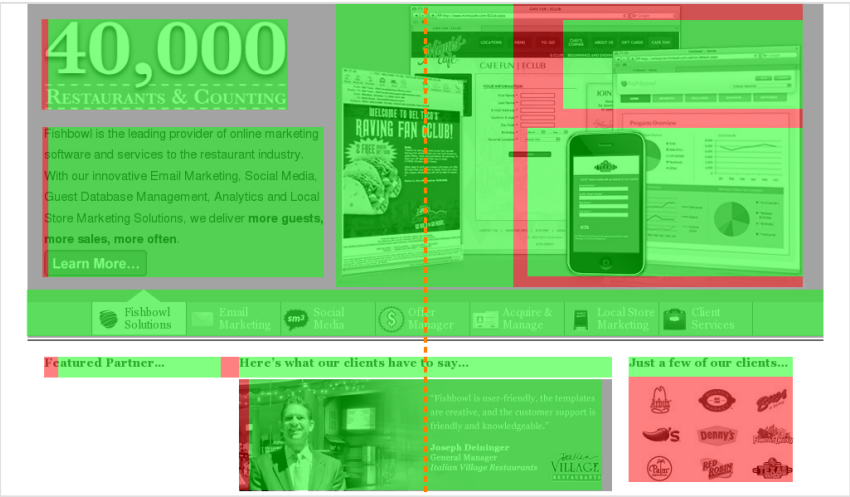


Figure 15. Vertical symmetry. The green blocks have a pair across the axis (orange, dotted); the red blocks (or block parts) are not paired.

3.1.3.2 Website Evaluation

The majority of past work has addressed the computation of webpage aesthetics; little work has addressed the computation of website holistic aesthetics. This paper addresses such a gap using two steps: adaptation of the for-webpage computational method to websites and introduction of website-level factors in aesthetics estimation. Both activities rely on viewing the holistic aesthetics computation as an improvement of the webpage aesthetics computation, and then, identifying how holistic aesthetics differs from webpage aesthetics and accounting for the difference.

3.1.3.2.1 Method Adaptation

The first step, adaptation of the for-webpage method to websites, included combining the aesthetics of individual pages in a holistic score and considering the full-length webpages in aesthetics computation. Different pages could differently impact the holistic aesthetics (e.g., a homepage might be the most impactful, since the user often sees it first). Such differences might need to be accounted for, e.g., in a regression formula, which would weight and sum up the aesthetics scores of individual pages according to the page importance. However, such regression-based method would require having a well-defined taxonomy of page types and weights of different types, which were unavailable. We used averaging instead. Using the automatic for-webpage measures of aesthetics (Section 3.1.3.1), we calculated the aesthetics scores for five pages per website and took their mean as the website-level scores of aesthetics. These averaged scores were used to select stimuli for the validation study presented in this paper.

The study also considered full-length webpages in aesthetics computation, exploring the kind of webpage screenshots – full-page or top-screen – that should be used in it. Many of past for-webpage methods of aesthetics estimation used only the top-screen parts of webpages. Such approach appeared reasonable, considering that the user first sees the top-screen parts and pays more attention to them (Djamasbi et al., 2011). However, the below-top parts might matter too. If the analysis of full-page screenshots explains significantly more of user score variance, they should be used despite they take longer to process. To select stimuli for the validation study presented in this paper, we followed the procedure from Miniukovich & De Angeli (2015a) and only computed website aesthetics for the top-screen page parts. We repeated the computation using full-page screenshots, compared the two approaches (looking at how well computed scores correlated with users' scores), and present the comparison.

3.1.3.2.2 Website-Level Factors

The second step addressed the introduction of website-level factors in aesthetics estimation, and included measuring visual diversity and website-level conceptual complexity. Visual diversity described the visual change across the webpages of website and was measured with two methods, metric-based and mpeg7-based. The metric-based method leveraged the for-webpage measures of

aesthetics and should have accounted well for the between-page change in the perception and appreciation of webpages. The measurement consisted of four step. We first calculated all for-webpage measures for five webpages per website. We then calculated the mean and standard deviation of scores across the five pages. The standard deviations were normalized by the means (cf., *coefficients of variation* from (Ivory & Hearst, 2002), and combined in a single averaged score of diversity per website.

The mpeg7-based method (cf., (Miniukovich & De Angeli, 2015b) leveraged six MPEG-7 and MPEG-7-based global image descriptors and should have accounted well for the between-page change in page content, as the descriptors were specifically designed to summarize the content and spatial distribution of content. The descriptors come from the image-processing field and have a range of applications, including the automatic match and retrieval of images. A diversity-related application of descriptors – retrieving a visually diverse image set from an image database – was also considered in the image-processing field (van Leuken et al., 2009). The descriptor list included the scalable color (SCD), color layout (CLD), dominant color (DCD), color and edge directivity (CEDD), edge histogram (EHD), and fuzzy color and texture histogram (FCTH) descriptors (Lux & Chatzichristofis, 2008; Iakovidou et al., 2014; Chatzichristofis & Boutalis, 2008). For each of the five webpage screenshots per website, we calculated descriptor values. We then looked at the pairwise distances between the five descriptor values, using the standard L1-norm (SCD, CLD & EHD), Euclidean norm (DCD) or Tanimoto coefficients (FCTH & CEDD). Five descriptor values allowed calculating ten pairwise distances. The mean of ten distances represented the visual diversity of a website (for one descriptor). Different descriptors gave the scores of different magnitude, and therefore, they were scaled before averaging in a holistic metric of visual diversity.

Conceptual complexity – which differs from perceptual, vision-based complexity – could decrease website aesthetics (Reber et al., 2004; Nadkarni & Gupta, 2007) and should be accounted for in aesthetics computation. Several website aspects could represent the conceptual complexity of websites, but we chose to consider the size and structural complexity of website. Website size was measured as an average page length, number of pages per website, and average time to load a page. We sampled up to a thousand webpages per website and took their mean length as an estimate of website page length and their mean loading time as an estimate of load time. The number of pages indexed by Google Search was taken as a measure of website size. We then looked at the complexity of site structure tree (Ivory & Hearst, 2002), using the size and breadth of the tree (traversed up to the 4th level, homepage was the level 1) and number of links per page. These three measures of site-tree complexity strongly cross-correlated and we kept only the number of links per page for selecting the stimuli tested in the validation study.

3.1.4 Study Method

We validated the method of website aesthetics computation in a user study that linked computed aesthetics scores to reported aesthetics scores. Study participants rated the websites that we selected from a larger pool of websites based on the computed aesthetics. The participants reported the scores of immediate, deliberate and post-use aesthetics. The latter was assumed to reflect the holistic aesthetics, which is the main focus of the paper. The study collected the evaluations on several UX dimensions other than aesthetics – website usability, content quality, and brand attitude – and tracked user behavior (the number of visited and revisited pages, number of clicks, browsing time per webpage, and time per task). We only report the results on aesthetics in this paper; the other results contribute little to the paper objective and will be reported elsewhere. Nonetheless, the complete data collection procedure is described below.

3.1.4.1 Stimuli Selection

We considered only corporate websites. Unlike news or eCommerce websites, corporate websites often have lower number of pages, which makes page counting possible (cf. site-level measures from (Ivory & Hearst, 2002). In addition, considering websites from several, unrelated domains could increase the number of confounding factors (cf. van Schaik & Ling, 2009). We further restricted the domain to the websites of civil engineering companies operating in the London area, which suited our use scenario. Each website included home, services or projects, about us, contact us, and career pages. The inclusion of these types of pages was based on our observation of the most common pages of corporate websites.

We outsourced the search and selection of websites to the workers of a crowdsourcing platform[9]. Such approach involved over 25 people sampling websites, which decreased the impact of sampler idiosyncrasies and excluded the authors from sampling (as they may unwittingly select the "right" stimuli). The crowdworkers searched for and reported the URLs of webpages of five civil engineering websites. This task lasted around 20 minutes and was compensated with 1.75 US dollars. A crowdworker could participate only once. We screened the reported websites and filtered out the ones that did not have all required pages, were small parts of bigger international portals or belonged to companies without an office in London. This resulted in an initial pool of 57 websites. We then took top-only (1660×1050 pixels) and full-width (1660 pixel wide, unlimited length) screenshots of the five pages of websites, which resulted in 285 screenshots (PNG format, 24-bit per pixel). To automatize screenshot taking, we developed an extension for the Mozilla Firefox v33.0 browser.

From the pool of 57 websites, we selected 15 high-aesthetics and 15 low-aesthetics websites. To categorize websites as high- or low-aesthetics, we used the algorithms and regression formulae from Miniukovich & De Angeli (2015a). We first computed aesthetics scores for the 285 screenshots (top-screen parts of pages, five page screenshots per each of 57 websites). We then took the means of the five screenshots as the aesthetics scores of websites. The high- and low-aesthetics groups included the websites from the left and right extremes of the 57-website aesthetics score distribution. A difference in user appreciation of two groups would demonstrate that the metrics indeed discriminated between appealing and unappealing websites. The final sample contained websites of varying appeal and complexity (Figure 16).

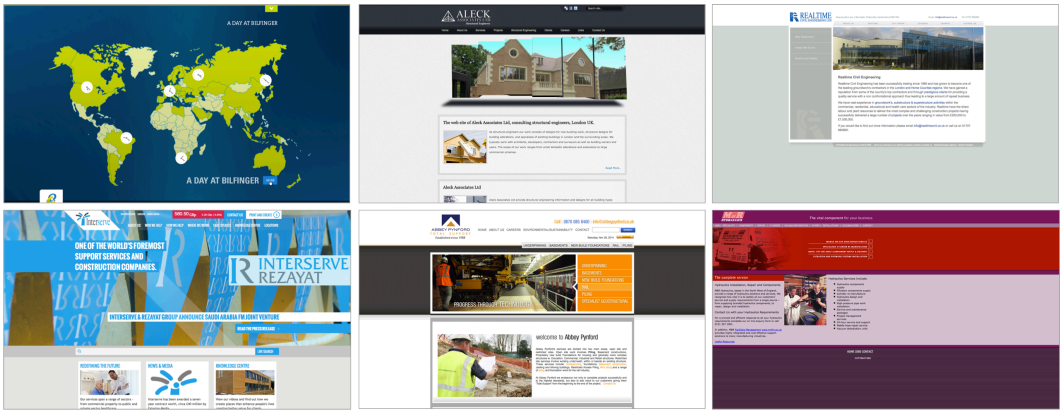


Figure 16. Examples of webpages used in the study, ranging from more complex (on the left) to less complex (on the right).

3.1.4.2 Experimental Design

A within-subjects design was used in the study, with website aesthetics as an independent variable manipulated at two levels: low and high. The manipulation was based on the *computed* scores of website aesthetics. The dependent variables included the *reported* scores of immediate, deliberate, and post-use aesthetics collected in user-evaluation study.

3.1.4.3 Participants

A class of computer science undergraduates participated in the study (N = 45, 8 female; mean age = 21.8 years, SD = 1.9; 2 computers crashed during the study, which brought N down to 43 participants). All participants reported to use the Web more than five hours a week. On a 5-point scale, participants reported to have adequate English proficiency (mean = 3.2, SD = 1.04) and moderate familiarity with GUI design (mean = 2.5, SD = .77). Around half of participants (23 out of 43) planned to look for an internship position in the near future.

[9] https://microworkers.com

3.1.4.4 Procedure

The first step included filling out a demographic questionnaire. Then, participants rated their immediate aesthetics impression of 20 webpages presented in a random order. We followed the procedure from Miniukovich & De Angeli (2015a) but used 200ms presentation durations instead of 150ms. We assumed longer presentation time should have increased the reliability of scores, which may be lower in group sessions due to more distractors than in individual sessions (Miniukovich & De Angeli, 2014b). The 200ms presentation time did not allow for scrolling – only the top-screen parts of pages were flashed on screens.

Then, the participants rated their deliberate-first aesthetics impression of the same 20 webpages presented in a different random order (Figure 17A). At this step, we used 4s presentation times: participants looked at screenshots for at least 4s and could scroll them down if the screenshots were longer than one screen.

Next, the participants read the use scenario, which asked them to imagine they were to select a company for an internship. The participants then proceeded to a Google Search-like interface featuring titles and short descriptive text snippets for four websites, Figure 17B. They opened and explored their four websites the way they deemed suitable during at least five minutes. The participants rearranged the order of websites in the search-result list, drag&droping the snippets of descriptive text (Figure 17B, right column) and continued to the information retrieval tasks. At this step (Figure 17C-D), the same four websites (but in a different random order) were searched, one after another, for required information and rated immediately after the search on four parameters (aesthetics, usability, content quality and company favorability). The order of questions was randomized.

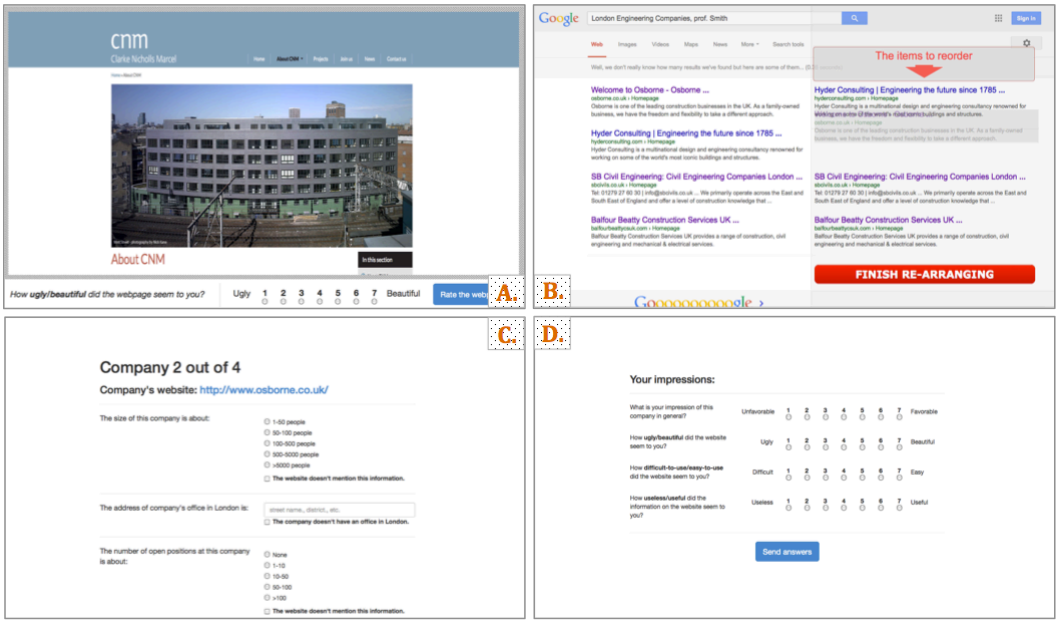


Figure 17. Data collection procedure, A) rating aesthetics of screenshots, B) company ranking (via drag&drop in the list on the right), and C-D) performing tasks on and rating websites.

3.1.4.4.1 Tasks and Use Scenario

We used two types of tasks: free browsing and information retrieval. The former was based on a use scenario and included participants freely going from one website to another and forming a holistic impression of the website and company. The exact wording of the use scenario was as follows. "*You are an engineering student and are looking for a 4-month internship position – doing an internship in a good place outside Italy would be a huge plus on your CV. Getting in such a place is hard. Your thesis supervisor worked in London for a while and has agreed to give you a recommendation letter to a few companies there. Googling 'London Engineering Companies, prof. Smith' retrieved a list of websites. However, you can only email one company at a time – be extra careful, consider the websites and rearrange the list, so the companies you want to get in the most are on the top.*"

In the information-retrieval task, participants searched for three specific pieces of information, namely, the size of website-related company, number of open job position in the company, and the address of company's office in London. Some of the websites did not feature the required information – we added an option "*The website doesn't mention this information*" as a possible answer to all questions. The participants rated each website right after answering the questions. This let us collect the ratings that reflected participants' most accurate post-use impression.

3.1.4.5 Instrument

We collected the scores of immediate, deliberate, and post-use aesthetics impressions, using two semantic differential items on a 7-point scale. One-item measures have been used before (Braddy et al., 2008; Sonderegger & Sauer, 2010; Tuch et al., 2009; 2012; Lindgaard et al., 2006). The exact wording of question was "*How ugly/beautiful did the webpage seem to you?*" for the first impression and "*How ugly/beautiful did the website seem to you?*" for the post-use impression.

3.1.4.6 Apparatus

Data collection took place in a group session in a classroom with computers (monitor resolution was 1680×1050 pixels) and lasted approximately 35 minutes. To present stimuli, collect user scores and track user browsing behavior, we developed an extension to Mozilla Firefox v33.0. After reading instructions and signing a consent form, participants opened a copy of Firefox with the extension pre-installed. Each participant was assigned a unique ID associated with a random selection of stimuli (four websites – two appealing and two unappealing– and 20 corresponding webpage screenshots).

3.1.5 Result

The study explored the computation of holistic aesthetics, which included two main activities: the adaptation of the for-webpage method of aesthetics computation to websites and validation of new website-level factors of aesthetics. The study also explored three additional phenomena related to the *computation* of holistic aesthetics. The phenomena included the impact of different webpages on the holistic aesthetics, comparison of top-screen and full-page screenshots in aesthetics computation, and evolution of aesthetics impression from immediate (200ms) to post-use impression (Thielsch et al., 2013).

3.1.5.1 Method Validation

If our computational method for websites could predict user aesthetics scores, the method validity would receive support. We considered three kinds of user aesthetics scores – immediate (200ms), deliberate (4s), and post-use – as dependent variables (DVs), and computed aesthetics category as a within-subjects factor (the algorithms assigned each website to either high or low aesthetics category). One of the DVs, deliberate (4s) aesthetics, was considered redundant due to its strong correlation with the other DVs, 200ms aesthetics ($r = .75$, $p < .001$) and post-use aesthetics ($r = .71$, $p < .001$). We followed the recommendation to exclude the redundant DV from the analysis (Tabachnick & Fidell, 2007). The correlation between the other two DVs (200ms and post-use aesthetics) was acceptable, $r = .59$ ($p < .001$). A total of 168 datapoints was reduced to 165 when we excluded multivariate outliers.

The immediate and post-use aesthetics were analyzed as two DVs in a MANOVA (a multivariate analysis of variance). The computed aesthetics factor served as an independent variable (IV). The aesthetics factor affected the DVs significantly, $F(2,162) = 7.86$, $p < .001$, partial $\eta^2 = .09$. The partial eta-squared demonstrates the amount of variance in the DVs that the binary factor could explain. The results thus show that an average participant evaluation reflected the computational evaluation: low-aesthetics websites were perceived as less appealing that high-aesthetics ones.

After testing both DVs in combination, we tested in isolation the variable of post-use aesthetics – this variable reflected holistic aesthetics impression, which this paper targeted. We performed an unpaired t test, with post-use aesthetics as a DV and computed aesthetics as an IV. The factor affected the DV significantly ($t(157) = 2.50$, $p < .05$). The difference in post-use aesthetics between the low- and high-aesthetics groups was diff. = .60 (low-aesthetics mean = 3.83; high-aesthetics mean = 4.43).

The MANOVA and t test suggested that the method could indeed differentiate the low-aesthetics websites from high-aesthetics websites. We then tested *how well* the method could predict holistic aesthetics and calculated Pearson's correlation between computed scores and post-use scores. The unit of analysis was a website. After inspecting the plot of computed-reported scores and removing an outlier (Figure 18), the correlation reached significance, $r(27) = .49$, $p < .01$.

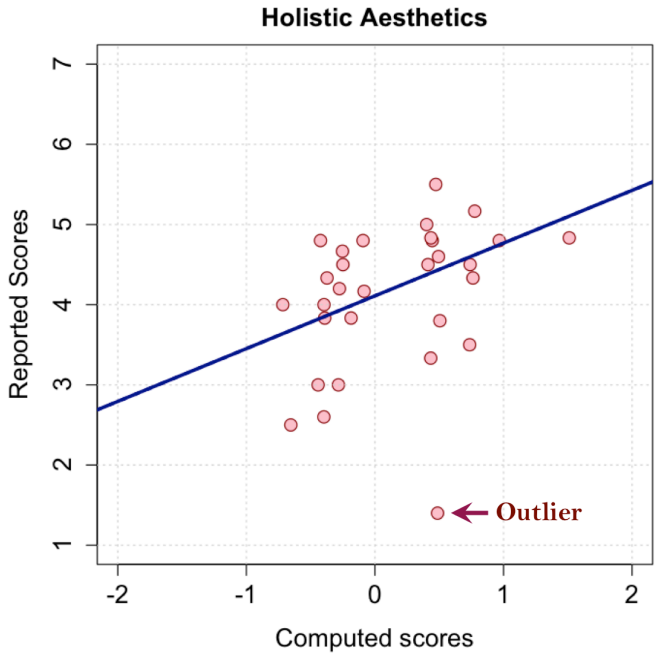


Figure 18. Plot of computed against reported scores of website aesthetics, with a regression line.

3.1.5.2 Website-Level Factors

We considered website conceptual complexity and visual diversity as they could affect the holistic aesthetics impression. Table 22 lists four measures used to estimate complexity; the distributions of sizes, number of links, and loading times (Table 22) were strongly positively skewed – we log-normalized these estimates to use them in further analyses. Out of four measures (Table 22), only log-normalized page counts (as indexed by Google Search) correlated significantly with website aesthetics ($r(28) = .56$, $p < .01$).

To estimate diversity, we used two methods: metric-based method, leveraging the means of variance in the scores of our eight measures for each website, and mpeg7-based method, leveraging the mean pairwise distance between the values of MPEG-7 image descriptors. Both methods generated the estimates that correlated with the reported post-use aesthetics. However, the mpeg7-based estimates correlated with aesthetics stronger ($r(28) = .51$; $p < .01$) than the metric-based estimates ($r(28) = .37$; $p < .05$).

		n	mean	sd	min	max	skew
B	Number of pages by Google Search	30	5300.10	22589.98	8.00	124000.00	4.83
C	Page length in screens* (1 screen = 950 pixels)	30	1.76	.75	.76	4.28	1.44
D	Links per page**	30	64.57	62.54	11.62	295.94	2.15
E	Loading time per page (in seconds) **	30	1.05	1.03	.19	4.94	1.98

* averaged from up to 1000 pages per site; ** averaged from all pages down to the 3rd level from a homepage.

Table 22. Descriptive statistics of automatically extracted meta-information about the websites.

We then tested if visual diversity (computed with the mpeg7-based method) and website size (log-normalized page counts) could explain extra variance in the holistic aesthetics, on top of the variance that the screenshot-based, 200ms aesthetics ratings could explain. We entered the two measures and 200ms aesthetics in a series of linear regression models (Table 23). The results showed both diversity and size to explain extra variance in holistic aesthetics, after the 200ms-based variance was controlled for (Models 2 and 3, Table 23). However, the explained extra variance was not large, around of 2-3% increase in model fit, R^2 (if Model 1 is considered a baseline model, $R^2 = .35$, and Model 4 an improved model, $R^2 = .38$).

IV	Model 1, β scores	Model 2, β scores	Model 3, β scores	Model 4, β scores
200ms aesthetics	.59***	.53***	.55***	.51***
Visual diversity	--	.17*	--	.12 (p = .11)
Website size	--	--	.15*	.11 (p = .14)
R^2 (R^2 adj.)	.35 (.35)	.38 (.37)	.37 (.37)	.38 (.37)

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

Table 23. The comparison of linear models of holistic aesthetics (DV).

3.1.5.3 Webpage Selection

Different webpage types could be used to compute a website aesthetics score. Homepages appeared to be the best webpage type to use in the computation: homepage-based computed scores correlated the strongest with the post-use scores (Table 24). To check if such homepage ability to predict the whole-website aesthetics stemmed from participants associating homepages and websites, we reviewed the per-page-type correlations between holistic scores, and 200ms and 4s aesthetics scores. The review did not indicate a page type that would stand out, which suggested that participants rated consistently all pages of the same website. A review of Cronbach's alphas further supported such hypothesis, revealing the high consistency of aesthetics within websites: the aesthetics scores of different page types could be treated as the items of a scale (5 page types, 200ms score alpha = .92, 4s score alpha = .95).

	Post-use aesthetics	
Page Types	r(28)	p
Homepage	.52	< .01
Contact us	.27	.14
Careers	.12	.53
Products & projects	.22	.24
About us	.24	.20
All 5 mean	.33	.07

Table 24. Pearson's correlations of computed aesthetics scores (separately for each page type) with post-use website aesthetics scores.

3.1.5.4 Full-Page Computation

We explored if relying on the full-length pages instead of top-screen page parts could have improved the accuracy of the method. During the 4s exposure part of study, participants could scroll down some screenshots (others represented pages too short for scrolling) and spent 37% of time looking at the page parts that required scrolling down to become visible (SD = .14, skew = -.39, min = .09, max = .64). This observation suggested full-page screenshots might indeed predict holistic aesthetics better than top-screen screenshots. We then computed the scores of the for-webpage measures (Miniukovich & De Angeli, 2015a) for both top-screen and full-length screenshots, and compared these two score sets. The scores of both sets correlated strongly with each other, with Pearson's r ranging from .71 to .97 (all $p < .001$; $df = 124$; mean $r = .84$). When analyzed in a linear regression, the full-page scores indeed predicted webpage aesthetics (best-fit model $R^2 = .41$; 95% conf. interv. = .32 to .56; $p < .001$; AIC = 292.5) better than the top-screen screenshots (best fit model $R^2 = .35$; 95% conf. interv. = .24 to .50; $p < .001$; AIC = 294.6). However, the largely overlapping R^2 95% confidence intervals and small (less than 2.0) difference in AICs (Akaike's Information Criterion describes the overall model fit penalized for each additional predictor) suggested the difference between the two models might be small.

3.1.5.5 Impression Evolution

User impression evolves from immediate through deliberate to post-use types (Thielsch et al., 2013). We collected the user scores of all three types of impression. The first two types included rating 20 webpages per participant after viewing them for 200ms (immediate aesthetics) and 4s (deliberate aesthetics). The latter type included rating 4 websites per participant after finishing tasks on the websites (post-use aesthetics). We aggregated the 200ms and 4s webpage scores in website scores (we had 5 webpages per website), and correlated all three types of aesthetics scores with each other (Figure 19). Deliberate aesthetics correlated strongly with both immediate ($r(163) = .75$, $p < .001$, 95% conf. interval = .67 to .81), and post-use aesthetics ($r(163) = .71$, $p < .001$, 95% conf. interval = .63 to .78); immediate and post-use aesthetics correlated less strongly ($r(163) = .59$, $p < .001$, 95% conf. interval = .48 to .68). The comparison of correlation magnitudes (using Fisher's z transformation) showed that $r = .59$ and $r = .75$ differed significantly ($z = 2.66$, $p < .01$), whereas $r = .59$ and $r = .71$ differed marginally significantly ($z = 1.89$, $p = .06$). Such differences in the correlations might suggest that immediate impression carried over in the reasoning-based post-use impression, and deliberate impression functioned as an intermediary between immediate and post-use impression.

3.1.6 Discussion

The study explored a method for the computational evaluation of website aesthetics. Several implications follow from the study and may help the researchers working on the computation of holistic website aesthetics.

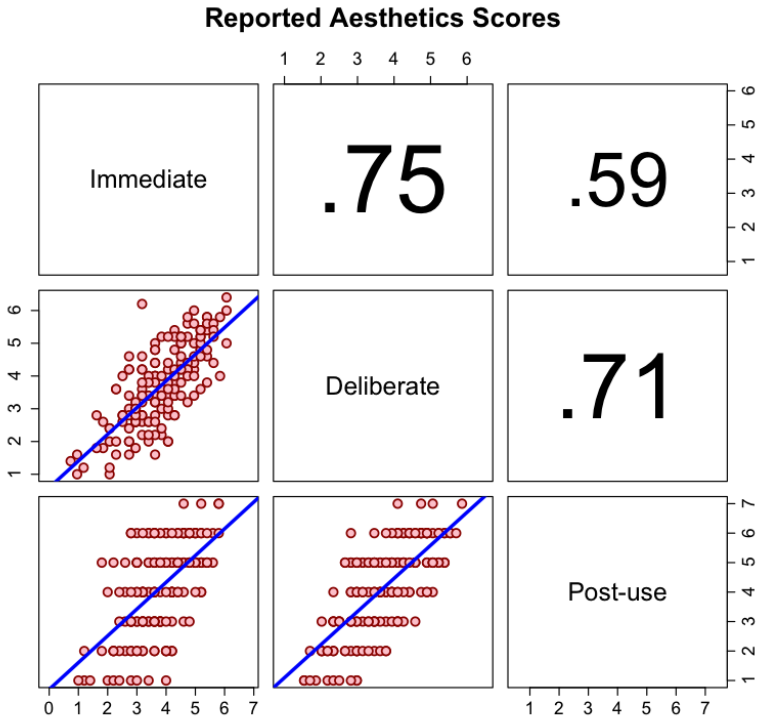


Figure 19. Plots and Pearson's correlations for the immediate, deliberate, and post-use aesthetics scores.

3.1.6.1 Method

The results suggested that the screenshot-based method could be used to estimate website holistic aesthetics. The method distinguished the websites of high and low aesthetics (with the magnitude of diff. = .60 on a 1 to 7 scale), and produced scores that correlated significantly with users' scores. Most past HCI studies on computational aesthetics modeled the aesthetics of webpages or webpage screenshots, and whether their method would function on websites remained unclear. But, the user

interacts with websites, not just views webpages: studies on computational aesthetics should aspire to model the post-interaction aesthetics, not just the viewing-based aesthetics. This paper explored the post-interaction, holistic aesthetics and demonstrated that the used method could model it.

3.1.6.2 Website-Level Factors

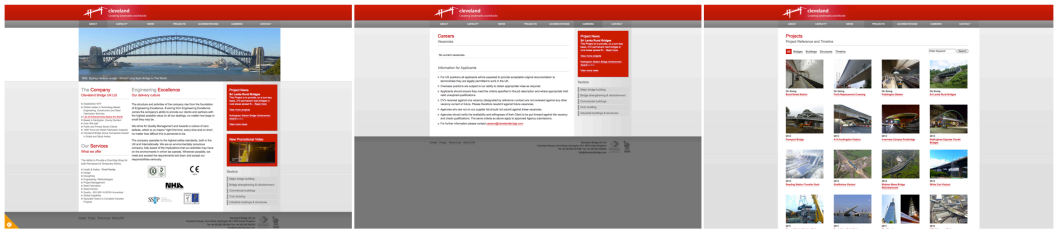
Automatically evaluating websites – rather than webpages – required accounting for website-level factors as such factors could impact the holistic website impression. The paper explored two such factors: cross-page visual diversity and website conceptual complexity. We used the mpeg7-based and metric-based methods to estimate visual diversity. The mpeg7-based estimates correlated stronger with the post-use aesthetics than the metric-based estimates. The substantial correlation between website aesthetics and visual diversity ($r(28) = .51$; $p < .01$) suggested that visual diversity should be included in the models of holistic aesthetics computation.

The regression analysis also suggested that visual diversity was a unique aspect of holistic aesthetics, different from visual complexity: when added in a regression model as an independent variable, visual diversity added up to the explained variance in the holistic aesthetics scores, on top of the variance that the complexity-based 200ms aesthetics explained (Table 23, model 3). Such analysis and result corroborated the proposal to include diversity in aesthetics models (Moshagen & Thielsch, 2010).

We initially expected that conceptual website complexity could decrease aesthetics (Reber et al., 2004; Nadkarni & Gupta, 2007). However, the hypothesis was not supported. We used four variables to estimate conceptual complexity (Table 22), and only website size (the log-normalized number of pages) correlated with website aesthetics. The unexpected direction of the correlation might suggest that the size-aesthetics relationship was indirect – the companies with bigger sites might have simply invested more in their image, their websites, and by extension, their website design. Such result might also suggest that conceptual complexity did not impact the evaluation of aesthetics. E.g., the user may have focused on the immediately visible part of website (visual complexity) and did not bother with the parts of website that required scrolling or clicking links to be seen (conceptual complexity).

3.1.6.3 Webpage Selection

Different webpages could have influenced differently the holistic aesthetics of websites. However, the results indicated that different pages had a similar impact on the website holistic aesthetics, as the different-page aesthetics scores stayed consistent within a website (Cronbach's alpha = .95 for deliberate aesthetics). Such consistency indicated that a single page per website could represent well website aesthetics. Further analysis indicated that homepages should serve as such representative webpage: the homepage-based computed scores of aesthetics correlated with the reported scores stronger than did the scores based on the other pages (Table 24). This could occur because homepages often contained a well-balanced mix of text and images (Figure 20), unlike other pages, which often contained only text (e.g., career pages) or only images (e.g., our-project pages). We conclude that using only homepages in webpage-level aesthetics computation might be acceptable, as a website always has a homepage and homepages often represent well the overall style of website.



A - Homepage B – Career page C – Our Projects page

Figure 20. An example of different pages: homepage (mix of text and images), career page (only text) and our-project page (mostly images).

Different aesthetics criteria might apply to different webpages. Such assumption could only be tested if a well-defined webpage taxonomy existed. Past research (Chen & Choi, 2008) attempted to suggest the taxonomies of page types, but may have failed due to confusing a webpage genre (e.g.,

"*information search page*") and website genre (e.g., "*online shopping*"). We introduced a taxonomy for the webpages of corporate websites, which consisted of five webpage types most common on the corporate websites. Despite these five webpages were most common, we still had to reject many initial-sample websites because they did not have all 5 webpages, e.g., if the contact-us and about-us pages were combined in one page. Future research should develop a robust and valid taxonomy for webpages.

3.1.6.4 Full-Page Computation

We explored the impact of below-the-top-screen parts of pages on website appreciation. Past work suggested the user mainly focused on the top-screen parts (Djamasbi et al., 2011), whereas our observations did not fully support this finding: our participants spent around 37% of time looking below the top screen (while rating deliberate, 4s aesthetics). The use of full-page screenshots – instead of top-screen screenshots – increased the predictive ability of the automatic aesthetics measures, but not substantially. The automatic-measure scores based on full-page and top-screen screenshots correlated strongly (avg. $r = .84$, $p < .001$), which could explain the non-substantial increase in predictive ability. Given such a slight benefit of using full-page screenshots, we conclude that top-screen webpage screenshots could be used to compute aesthetics. However, future work should this conclusion (Figure 21).

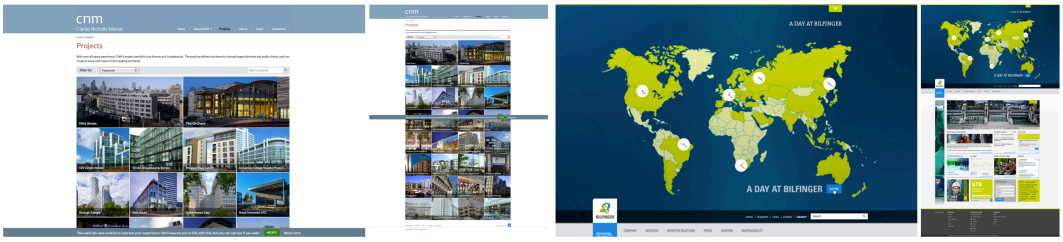


Figure 21. An example of top-screen page part featuring the same ratio of text to image as the whole page, and thus, representing well the whole webpage (left), and example of the opposite (right).

3.1.6.5 Impression Evolution

Collecting the scores of post-use impression requires participants to not just view, but actively browse websites. Such requirement limits the number of websites a participant could evaluate and increases researcher's effort to collect the amount of data needed for a model development. If the scores of immediate impression could be used, this would decrease the effort. Our analysis revealed strong correlations between reported immediate and deliberate aesthetics, and post-use aesthetics (Figure 19), which suggested that users' immediate impression carried over into the holistic judgments of aesthetics, as expected (Deng & Poole, 2010; Thielsch et al., 2013). We concluded immediate aesthetics could stand in for post-use aesthetics if the ecological validity of user aesthetics scores were not crucial.

3.1.7 Conclusion and Future Work

The paper has showed that the automatic estimation of website holistic aesthetics is feasible, which encourages further work on computational aesthetics in HCI. We sampled and processed multiple websites, re-created a relatively realistic use context in the user study, and analyzed website-level factors of holistic aesthetics (between-page visual diversity and website conceptual complexity). Our aesthetics-estimation method also used relatively few computational features and simple regression formulae, which could let designers interpret the estimates of each feature and facilitate their use to aid design – the method would quickly provide feedback on each feature of designs (Rosenholtz et al., 2011).

The results of the study were positive, however, future work should extend and confirm them while addressing several issues. First, we have not considered within-page dynamics, such as animation (e.g., rotating ad banners) and dynamic visualizations (e.g., unfolding navigation menus), cf., (Tractinsky et al., 2011; Huhtala et al., 2011). The user might like user-triggered GUI changes (e.g., menu unfolding triggered by a click), but dislike automatic GUI changes (e.g., rotating banners). A future study should investigate user preferences for animation types and speed. Second,

the cross-page visual diversity of websites should be explored further. Our results showed that diversity is related to the holistic aesthetics. However, defining the nature of such a relationship (e.g., a causal relationship), and integrating the diversity in the models of UX remain a challenge. Finally, future research could further extend our aesthetics computation method and consider a number of GUI aspects that HCI and psychology research indicated to increase liking, e.g., color equalization (Rizzi & McCann, 2007), which improves readability, and thus, aesthetics. Or, future research could exploit the peak-shift effect (liking of stimuli that differ slightly from the prototype, (Ramachandran & Hirstein, 1999) by collecting user expectations of GUI design on a large scale (Kumar et al., 2013), calculating the visual distance between a design and design expectation, and correlating the distance with aesthetics ratings.

3.2 Visual Diversity and User Interface Quality

Live graphical user interfaces (GUIs) do change responding to user actions, unlike GUI screenshots, which are often used in studies. The user experiences and is affected by transitions between the layouts (e.g., webpages or mobile app screens) of interactive systems. Such transitions affect the overall impression of system quality and should be accounted for by any model or computational method estimating the quality and claiming high ecological validity. However, the recent efforts aspiring to predict GUI quality computationally have only relied on homepages or home screens of apps, or their screenshots. The dynamics of GUI – GUI change across pages and layouts, or shorter, visual diversity – have been given little attention. Here we present an initial exploration of GUI visual diversity. In three studies, we demonstrate that a) GUI diversity can be measured computationally; b) GUI diversity correlates with GUI aesthetics impression and other, more high-level GUI-preference constructs; and c) GUI diversity matters in both website and mobile app contexts. We believe the concept of GUI visual diversity deserves further studies.

3.2.1 INTRODUCTION

Users interact with the whole of website or mobile app, not with distinct pages or app screens – the transitions between pages or screens also influence users' post-use impressions. A part of human-computer interaction (HCI) studies – mainly concerned with the high-level concepts of GUI quality (e.g., usability, meaningfulness or aesthetics) – used live websites and apps, and did account for the transition component of user experience (UX). However, those were questionnaire-based studies exploring the interplay of various GUI quality constructs (e.g., Lee & Koubek, 2010; Hartmann et al., 2007; Thielsch et al., 2013; Tractinsky et al., 2006). Computation-based approaches to GUI quality have seen a recent surge of interest, however the studies have largely ignored everything beyond homepages (Miniukovich & De Angeli, 2014a; 2014b; Reinecke et al., 2013; Reinecke & Gajos, 2014; Wu et al., 2013; Zheng et al., 2009). Even the few studies (Miniukovich & De Angeli, 2015a) that did consider other pages still treated the pages as distinct, unrelated inputs to their algorithms. The studies have not systematically addressed the possible impact of visual difference across same-website pages.

The automatic systems of GUI quality assessment should evolve to account for the holistic user impression, not only the impression of homepage screenshots. Such impression combines the impact of different system layouts (website pages or app screens, cf. the discussion in (van Schaik & Ling, 2009)), the impact of dynamic visualization and animation (Tractinsky, 2013; Tractinsky et al., 2011; Leuthold et al., 2011), and the impact of between-layout visual consistency (van der Geest & Loorbach, 2005, the opposite of visual diversity). In this paper, we concentrate on the issue of visual diversity. The pages of a website or screens of an app differ visually; the magnitude of difference may affect the user impression of a system. Past research argued the effect of diversity might be detrimental (e.g., higher cognitive load to process each new, visually-different page, cf., (Robertson et al., 2009; Beck et al., 2009)) or beneficial (e.g., helping to distinguish between different content areas of a website, cf. van der Geest & Loorbach, 2005). However, the definitive conclusion has not been drawn (cf., Moore et al., 2005).

This paper presents an exploratory study of layout (website and mobile app) visual diversity and diversity computation. We define visual diversity as the amount of possible visible change across the webpages of site (cf., van der Geest & Loorbach, 2005) or layouts of mobile app. The automatic estimation of visual diversity has not been practiced in HCI – we estimated the diversity using algorithms from the image retrieval research area (Lux & Chatzichristofis, 2008). In three studies, we processed three separate sets of stimuli, two website sets and one mobile app set. Each consecutive step included the user scores of higher ecological validity: in the first study, participants only rated the immediate aesthetics of webpage screenshots; in the second study, participants interacted with and rated live websites; in the last study, real users rated existing mobile apps (we collected the data from Google Play Store). Overall, the results showed that users associated higher visual diversity with higher website aesthetics and app popularity, with few exceptions. Website usability and content quality did not correlate with visual diversity. The remainder of the paper includes the description of related HCI research and visual diversity metrics that we used. The description of the

studies and discussion of results follow. We conclude by drawing the implications for future research, and call for more studies.

3.2.2 RELATED WORK

Researchers (Grudin, 1989; van der Geest & Loorbach, 2005) distinguished three types of interface consistency. (Past literature tended to mention the term consistency, which we view as the opposite of diversity.) First type, internal consistency, accounts for the consistency of GUI visual and physical layout within a tool – consistency in command naming, types of used UI elements and their placement within a layout, and color, shape and size of the elements. Second type, external consistency, accounts for the use of existing GUI conventions – conventions describing the aspect mentioned in the sentence above – and is supposed to help new users learn an interface. A recent study (Roth et al., 2010), for example, demonstrated that users expect to find certain webpage elements (e.g., sign in/login fields) at certain screen locations (e.g., the top-right corner). Violating such expectations might increase user frustrations or provoke curiosity. The last type, real-world consistency, accounts for the resemblance between interface elements and everyday objects. Familiar metaphors from the real world are supposed to increase technology acceptance.

In this paper, we explore the first type of consistency, internal consistency. We further narrow down the exploration domain to the visual consistency across static states of software systems, i.e., across webpages of websites and mobile app screens (often called *activities* in the Android development). We leave out from consideration the visual (in)consistency occurring within a single GUI state, e.g., due to dynamic visualization (unfolding menus) or animation. The initial explorations of such GUI dynamics – e.g., studying people's aesthetic perceptions of in-vehicle animations (Tractinsky et al., 2011) or linking the types of animation to the perception of animation duration (Huhtala et al., 2011) – have recently advanced our understanding of GUI dynamics effects on GUI quality. However, a separate comprehensive study needs to advance these works, which we consider our potential next step. Here, we focus on the between-screen visual consistency, measuring it computationally, and linking it to higher-level GUI quality constructs.

3.2.2.1 GUI Quality

A deeper understanding of the concept of visual diversity would require fitting it within the existing GUI quality frameworks. Several frameworks have been proposed; all of them differ substantially. Hassenzahl (2005) distinguished between pragmatic and hedonic system qualities. Both types of qualities stem from lower-level design features (such as, system functionality, presentation or content) and result in such outcomes as system appeal or post-use satisfaction. Pragmatic qualities regard interactive systems as tools helping to reach functional goals. Hedonic qualities regard interactive systems as bringing meaning in users' lives, reflecting users' personalities or stimulating playful exploration. Both pragmatic and hedonic qualities have been further considered within the context of interactive system *goodness* and *beauty*: goodness depended on both pragmatic and hedonic, whereas beauty solely depended on hedonic (van Schaik & Ling, 2009; Hassenzahl, 2004).

Kim and Fesenmaier (2008) used a different approach to website quality. They described two types of design factors: hygiene and potential factors. The former included website informativeness and usability, and were considered a necessary, but not sufficient component of website success. The latter, potential factors, included website credibility, inspiration, involvement and reciprocity, and were considered the "extra" necessary for a website to stand out. The authors (Kim & Fesenmaier, 2008) tested their model on tourism websites, and showed inspiration and usability to be the primary drivers of website first impression.

Rafaeli et al. (2004) proposed and tested a three-dimension model of emotional response to a physical artifact. The first dimension – instrumentality – represented the effectiveness and efficiency of the artifact as a tool. Notably, poor performance of the artifact would almost always lead to frustration, whereas satisfactory performance would be barely noticed. The second dimension – aesthetics – represented the sensual impact of the artifact, without cognitive mediation. Rafaeli et al.'s participants often strongly (dis)liked artifact appearance without giving a reason for the (dis)liking. The last dimension – symbolism – represented the associations triggered by the artifact. Tractinsky et al. (2006) later tested the model within HCI using multi-item scales and found the three dimensions to be relevant and independent.

3.2.2.2 Visual Diversity

Visual diversity has been rarely considered in HCI models of GUI quality, despite it may well be connected with many of the GUI quality components, e.g., with three most-studied components (cf., Hartmann et al., 2007; Thielsch et al., 2013; Hartmann et al., 2008): usability, aesthetics, and content quality. Higher diversity could hamper user performance (lower usability); lower diversity could make a GUI seem consistent (higher aesthetics), or – if pushed to an extreme – could make the GUI simplistic and primitive (lower aesthetics); finally, higher diversity could correspond to novel and diverse content (higher content quality). Contradictory accounts of visual diversity impact can be drawn from relevant literature. We present both cases: in favor of consistency and in favor of diversity.

3.2.2.2.1 Higher Consistency

Preserving users' mental map – the term is widely used in graph visualization literature, interchangeably with higher consistency or spatial stability – is considered a pillar of good information visualization (Robertson et al., 2009), even further, an aesthetic criterion for dynamic visualizations (Beck et al., 2009). User commit fewer errors and spend less time when operate on the dynamic graphs that preserve mental maps (Archambault & Purchase, 2013), which should increase their satisfaction and perceived usability. Nadkarni et al. (2007) considered visual consistency a component of perceived website complexity. Lower consistency implied higher complexity, leading to dissatisfaction. Nadkarni et al. (2007) asked participants to rate the similarity of webpage graphics and "information items" across pages, and even attempted to measure consistency[10] computationally, as the variation in the number of content types used on webpages (texts, graphics, video, audio, animation). The consistency, however, was not the focus of study and was combined with 12 other measures in an index of objective website complexity.

In a study of website menu types, Leuthold et al. (2011) observed participants to take fewer eye fixations and shorter times to accomplish tasks if all menu items were simultaneously visible on the screen. Structuring items in dynamic, unfolding menus had the opposite effect – performance dropped, and perceived effort and frustration soared. Visual inconsistency due to folding/unfolding menus might well account for a part of these negative effects. Van der Geest et al. (2005) studied user capability to recognize and explain visual inconsistency. They showed to participants interface elements drawn from different webpages, and asked them to group the elements and label the groups. If participants disagreed on a group to put an element in, this was considered a sign of visual inconsistency. Participants used six types of labels to describe within-group consistency: descriptive (e.g., *orange* or *horizontal*), associative (e.g., *scientific* or *tranquil*), function-related (e.g., *product information* elements), unlike-the-other (e.g., *a style of its own*), combination of the above, and uncategorized labels. Notably, the descriptive type of labels accounted for more than a half of all labels, with color being the most prominent visual cue (the other cues included background, font, illustration, grid/navigation and logo). Despite van der Geest et al. (2005) did not explicitly link visual inconsistency to performance or error rate, they still argued for the use of consistency in design. In particular, they suggested using the same color schema within a website, and similarly shaped, sized and placed GUI elements to convey within-page consistency.

3.2.2.2.2 Higher Diversity

Several cases in favor of higher visual diversity have been presented. Grudin (1989) admitted that, amid the prevailing calls to "strive for consistency" in design literature, interface consistency may appear excessive and limit the flexibility or adaptability of system. For example, keeping a GUI popup menu consistent supports the ease of learning but may conflict with the ease of use. A more efficient strategy could be placing the most expected-to-be-next item (as in the copy-paste two-step operation) on the top of menu. In another example that Grudin (1989) presented, applying the "*print*" operation to a folder should print the list of folder documents and their meta-information, rather than the documents themselves. Such logic would be consistent with system architecture, as understood by the system developers. However, the user would be more interested in printing out all documents in one click.

[10] Strictly speaking, Nadkarni et al. (2007) measured the consistency of webpage media types, not *visual* consistency.

Moore et al. (2005) explored the impact of advertisement-website consistency on recall, recognition and attitude. Their results (Moore et al., 2005, experiment 1) showed participants favored website-consistent ads, but remembered better website-inconsistent ads. A follow-up experiment (Moore et al., 2005, experiment 2) showed moderate inconsistency to be the most beneficial, both for ad attitude, and ad recall and recognition.

Finally, systems offering higher diversity may be viewed as more beautiful. Armstrong et al. (2008) argued that simple, regular, consistent objects could be merely perceived as pretty. To qualify as beautiful, the objects should offer the prospects of knowledge reshaping and expansion, i.e., novelty. Ramachandran et al. (1999) linked aesthetic pleasure to multiple successive steps of meaning discovery. Each step results in a small reward and stimulates the sense of pleasure. In their account, a completely familiar object would be seen as not worthy exploration and dull. Similarly, Kaplan (1973) argued humans preferred the environments they could make sense of (i.e., consistent with past experience), but also novel, challenging and uncertain. In such case, humans can integrate new knowledge within their existing cognitive map (Kaplan, 1973) and are evolutionary rewarded for that. Thus, offering some diversity in a GUI – without excessive departures from the familiar and consistent – may stimulate curiosity and be rewarding to experience.

3.2.2.3 Measures of Visual Diversity

Visual diversity could be estimated in several ways. Nadkarni et al. (2007) counted the number of media types (e.g., text, image or audio) used on pages, and then, estimated the variability in these numbers across pages of websites. Harper et al. (2013) split webpages in 300-pixel square blocks and counted the number of top-left corners of HTML elements falling in each block. This could be seen as a within-page variability measure, but could also be easily transformed in a between-page measure. Ivory et al. (2002) estimated webpage consistency with the Coefficients of Variation. They computed a range of page-level parameters (e.g., the number of used fonts or text colors); the standard deviations of the parameters normalized by their means were the variation coefficients. However, all aforementioned measures analyzed the underlying codes of GUIs (e.g., HTML of webpages), which did not reflect precisely what the user saw.

The screenshots of GUIs reflect precisely what the user sees. Analyzing the visual difference between them could reflect well the visual diversity within a website or app. Several such measures of visual difference have been tested in the image indexing and retrieval domain, and appeared to perform well on image search (Iakovidou et al., 2014) and image diversification tasks (van Leuken et al., 2009). The measures are based on the MPEG-7 and MPEG-7-like global image descriptors, namely scalable color (SC), color layout (CL), dominant color (DC), edge histogram (EH), color and edge directivity histogram (CEDD), and fuzzy color and texture histogram (FCTH) descriptors. SC, CL and DC solely describe image colors; EH solely describes image contours; CEDD and FCTH describe both image colors and edges.

- (SC) Scalable color describes the basic color distribution of image. First, image colors (in the HSV color space) are quantized into 256 bins, which are then compressed into 64 coefficients, using the Haar transformation.
- (CL) Color layout describes the global spatial distribution of image colors. An image is divided in 64 (8 × 8) blocks, and the main colors (in the YCbCr color space) of each block are determined. The discrete cosine transformation is then used to calculate 192 (64 blocks × 3 color channels) coefficients, which represent the image.
- (DC) Dominant colors describe the main, representative colors of image. All pixel colors of image are clustered in up to eight clusters. The descriptor includes the centers of clusters and proportion of image covered by each cluster.
- (EH) Edge histogram describes the spatial distribution of image contours and their directionality. An image is divided in 16 (4 × 4) blocks; for each block, five types of edges are detected (vertical, horizontal, 45- and 135-degree diagonal, and isotropic edges). This results in 80 (16 blocks × 5 edge types) coefficients per image.
- (CEDD) Color and edge directivity histogram describes both color and texture of image. An image is divided in a present number of rectangular blocks; each block is classified in one of 24 color clusters (i.e., a 24-item custom color palette is used) and one of 30 texture clusters (5 edge directionalities – similar to the EH descriptor above – and 6 thresholds are used). Lastly, color and texture histograms are combined in a single 54-item descriptor.

- (FCTH) Fuzzy color and texture histogram resembles CEDD, but uses the Haar wavelet Transform to cluster blocks by their texture and results in longer, 72-item descriptors.

3.2.3 Exploration of Visual Diversity

We computed website or mobile app visual diversity based on the pairwise comparison of several screenshots of the website or mobile app. The same procedure was used for all six aforementioned image descriptors. First, a descriptor is computed for a pair of screenshots. Then, a distance between the two screenshots is computed (we used the Tanimoto coefficients for calculating CEDD and FCTH distances, the L_1 norm for CL and EH distances, and the Euclidean norm for DC and SC distances). The operation is repeated for all possible pair combinations of screenshots of a website or mobile app (this result in n*(n-1) comparisons, where n is the number of screenshots). Then, for each descriptor, the mean over all pair distances is taken as the measure of visual diversity. Finally, the scaled means for each of the six descriptors was further averaged into a single, overall metric of visual diversity.

In three exploratory studies, we analyzed the visual diversity of three different sets of stimuli, two sets of website screenshots and one set of mobile app screenshots. The main requirement to the datasets was the inclusion of multiple screenshots per each website or mobile app. Each consecutive study involved a higher degree of ecological validity. We started from the premise that both GUI visual diversity and visual aesthetics might describe overarching GUI goodness, and thus, the two might correlate. We tested the premise in study 1.

3.2.3.1 Study 1

The dataset of study 1 was collected as a part of past exploration of website visual aesthetics (Miniukovich & De Angeli, 2015a, study 1). Using a 1-7 semantic differential scale, participants rated the immediate aesthetics (150ms exposure) of 300 webpage screenshots of 75 websites. All websites came from one of three genres (corporate, news or eCommerce) in equal parts, 25 websites per genre. Four webpage screenshots were taken for each website. The types of the four webpages were kept the same within a genre, e.g., home, about us, contact us, and our service pages for the corporate genre websites. The selection of such 75-website pool was outsourced to the workers of crowdsourcing platform – a relatively high number of selectors should have alleviated the impact of individual preferences on selection. Further details about the dataset can be found in (Miniukovich & De Angeli, 2015a).

At the time of study, we expected visual diversity to be linked to website aesthetics, e.g., through appealing graphics. Such graphics, if placed on different pages, could simultaneously correspond to higher aesthetics and higher diversity. We, thus, expected to observe the aesthetics-diversity link despite participants viewed and rated individual webpage screenshots shown on a screen, not live websites. We took the mean of four webpages per website as the aesthetics score of website. We then computed visual diversity scores for the websites and matched them against the aesthetics scores.

3.2.3.1.1 Results

We reviewed the histogram of the six visual diversity measures; the distribution of scores by the DC measure was positively skewed and we log-normalized it. The Cronbach's alpha of .82 suggested the 6 measures could be treated as sub-items of a scale – we filtered out 3*SD outliers, scaled and computed the average of scores of the six measures. The review of aesthetics-diversity plots for all six measures and their average suggested a positive linear relationship between the aesthetics and diversity scores, e.g., Figure 22. Visual diversity indeed correlated with website aesthetics, Table 25.

3.2.3.1.2 Discussion

The analysis revealed a positive correlation between website aesthetics and website between-page visual diversity (Table 25), which suggested that the two constructs were related. However, the dataset only included the scores of webpage aesthetics (no other GUI aspects were rated), which could not let us check if diversity related to aesthetics directly rather than indirectly. For example, website content quality could link the two concepts: high between-page diversity could correspond to the diversity of interesting content; high aesthetics could also coincide with interesting content, e.g., as the result of the halo effect (Hartmann et al., 2007). Other website GUI aspects should have been tested. A true test of between-page visual diversity would also require participants to interact with live websites and experience between-page transitions, rather than passively view webpages.

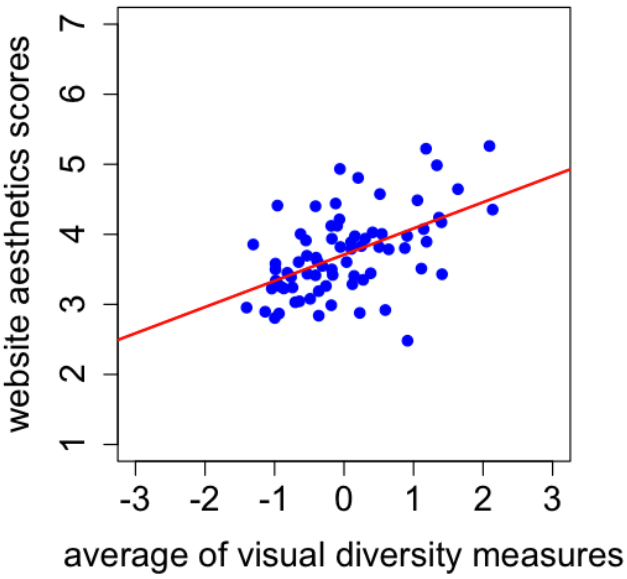


Figure 22. Plot of website aesthetics scores against the average of six diversity measures (scaled and centred).

Measures	Pearson's r
CEDD	.51***
CLD	.46**
EHD	.28*
SCD	.43***
FCTH	.48**
DCD	.35**
Average	.52***

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

Table 25. The Pearson's correlations between website aesthetics scores and scores of visual diversity measures, df varied 71 to 73 after 3*SD outliers were filtered out.

3.2.3.2 Study 2

Study 2 addressed the limitations of study 1: it involved rating several qualities (not only aesthetics) of live websites (not screenshots). Small conceptual overlapping across the GUI quality frameworks (see, section 2.1) could complicate choosing the constructs to investigate; many of the constructs are domain-specific, e.g., *inspiration* (Kim & Fesenmaier, 2008) was tailored specifically for travel websites. We chose the most-often studied components – website usability, aesthetics, and content quality – and company favorability (personal preferences for a company as a job place; each website belonged to a company). Such selection seemed to be a well-rounded representation of the frameworks.

Using live websites would ensure participants did experience visual diversity; collecting ratings of usability, aesthetics, content quality and company favorability right after the experience might tentatively suggest how visual diversity impacted participants. One, for example, might have expected higher diversity to be associated with more novel, interesting content, and thus, correspond to website aesthetics indirectly.

3.2.3.2.1 Data Collection

We collected the dataset of study 2 as a part of bigger study of website UX. Five webpage screenshots were taken for each of selected 30 websites of London-based civil-engineering companies. (All websites could be classified as corporate-type.) A relatively big number of engaged website

selectors – the workers of crowdsourcing platform – should have mitigated the selection bias due to individual preferences. The types of webpages were the same for all websites: home, about us, contact us, career, and our-projects pages. We used the screenshots to compute website visual diversity. The diversity measures were the same as in study 1.

After reading a briefing form and clarifying doubts, each participant (students, total N = 45, 8 female; mean age = 21.8 years, SD = 1.9) performed three information retrieval tasks on four live websites, one website at a time. The tasks included finding the address of company HQ, the number of employees, and number of open positions for interns. Right after tasks, a participant rated the four qualities of a website (usability, aesthetics, content quality, and company favorability) on a 1-7 semantic differential scale. Experimental sessions lasted 25 to 40 minutes; no fatigue complaints were reported. User ratings (five to six per website for each construct) were aggregated and matched against the computed scores of visual diversity.

3.2.3.2.2 Results

The review of six diversity measure histograms again showed that only the scores of DC measure were positively skewed. We, therefore, log-normalized them for the consecutive analyses. The Cronbach's alpha of .91 (n = 6) suggested the diversity measures could be treated as the sub-items of a scale. We filtered out 3*SD outliers, scaled and averaged the scores of six measures in the index measure of visual diversity (the index measure was used to select the examples, Table 25). The visual diversity measures correlated well with aesthetics scores, less well with company favorability scores, and did not correlate with usability and content quality scores, Table 26. If visual diversity and websites aesthetics were considered as predictors of company favorability, the impact of visual diversity on favorability would become non-significant (Table 27). This suggested the full mediation of diversity-favorability link by aesthetics.

Diversity measures	Aesthetics	Usability	Content Quality	Company Favorability
CEDD	.40*	.10	-.05	.28
CLD	.36	-.03	-.19	.21
EHD	.53**	.05	.03	.46*
SCD	.53**	.09	-.04	.31
FCTH	.51**	.12	.06	.39*
DCD	.46*	.37*	.11	.28
Average	.56**	.11	.01	.41*

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

Table 26. Pearson's correlation coefficients describing the link between visual diversity scores and post-use scores of website quality, df ranges 28 to 29 due to filtering 3*SD outliers.

3.2.3.2.3 Discussion

As in study 1, we observed a strong positive link between website visual diversity and website aesthetics (Table 26). Our expectation of this link being potentially mediated by content quality was not confirmed: website visual diversity was unrelated to the perceived website content quality (Table 26). Neither was visual diversity related to website perceived usability. Visual diversity did correlate with company favorability, however the effect appeared to be fully mediated by aesthetics (Table 27). Thus, we might suggest visual diversity exclusively impacted the beauty of websites; all impacts of diversity on website approach-avoidance tendencies – if discovered in later studies on, e.g., revisiting or friend recommendation – might stem from higher aesthetics.

3.2.3.3 Study *3*

The datasets of studies 1 and 2 resulted from in-lab experiments. One might argue that an in-lab UX differed from the real-world UX, and in-lab assessments and scores would change in a real-life use context. Study 3 addressed this concern: we scraped the real-user assessments of real-world apps from the Google Play website. Study 3 explored mobile apps instead of websites and used larger samples than those in study 1 and 2. This should have increased the reliability and generalizability of study 1 and study 2 claim that visual diversity was related to GUI quality.

	Model 1	Model 2	Model 3
Diversity	.03	.41*	--
Aesthetics	.67***	--	.69***

*** $p < .001$; * $p < .05$

Table 27. Beta coefficients for three linear regression models; company favorability scores are the outcome. A visual diversity impact is not significant in Model 1, i.e., appears to be fully mediated by aesthetics.

3.2.3.3.1 Data Collection

We sampled app screenshots, ratings and number of ratings from Google Play. Developers upload the screenshots of their apps on the Google Play website; whereas users rate and comment on the apps they have chosen to install. We assumed the ratings and per-app number of ratings to describe app popularity. Three app categories were considered: shopping (188 app), news and magazines (157 apps), and business (107 apps). We only sampled the apps that had five or more unprocessed, non-duplicate GUI screenshots uploaded on their description page (many app developers uploaded different-resolution screenshots of exactly the same GUI; others processed the screenshots, adding signs or "embedding" them in mobile devices and showing the photographs of the devices). We also excluded the apps that did not belong to the stated categories (e.g., games about business did not qualify as business apps) and avoided apps for tablets. We excluded the apps that had fewer than 50 ratings. Lastly, the purpose, functionality and layout of mobile apps varied widely even for the apps of same app genre. Thus, we could not sample same-purpose layouts for apps, as we did in study 1 and 2 for websites.

3.2.3.3.2 Results

A large number of sampled apps allowed us consider each category separately. The scores of DC measure were log-normalized, as before. The histograms of app rating numbers also showed a lognormal distribution; the numbers were log-normalized. The number of app ratings correlated with the ratings in all categories, shopping ($r(186) = .40$, $p < .001$), news and magazines ($r(155) = .41$, $p < .001$), and business ($r(105) = .44$, $p < .001$).

The Cronbach's alphas were high for all three categories, shopping (alpha = .75, n = 6), news (alpha .79, n = 6), and business (alpha = .80, n = 6). We combined the scores of the six diversity measures in the index diversity measures, again, after filtering out 3*SD outliers and scaling. None of the diversity measures correlated with app ratings in any category. However, the number of app ratings did correlate with some of the diversity measures, positively for the news and business categories, and negatively for the shopping category (Figure 23). Further pairwise unpaired t-tests revealed the differences in visual diversity between categories: the news apps were the most diverse (significant difference for all 6 measures); the shopping apps were more diverse than business apps (significant difference for the EH and SC measures). For each category, we additionally reviewed a plot of index diversity measure against the number of app ratings, with a Lowess (locally weighted scatterplot smoothing) curve fitted. No indication of non-linear relationships was observed.

3.2.3.3.3 Discussion

App visual diversity correlated with the number of times the app was rated or commented on, positively for news and business apps, and negatively for shopping apps (Table 28). We suspected the diversity within shopping apps could be extreme and cause a backlash (i.e., the negative correlation). However, the shopping apps were significantly less diverse than the news apps and only slightly more diverse than the business apps. Thus, other than diversity factors should have caused the inverse correlation. We might speculate the standards of goodness differ across categories. For example, shopping apps (Figure 24) might feature more flashy, graphical ads, which increase diversity, but surely worsen app impression. Also, the users of shopping apps might be much concerned with app security and credibility – the qualities associated with lower visual diversity (cf. (van der Geest & Loorbach, 2005)). On the other hand, the users of news apps (Figure 25) do not risk their money and fearlessly explore new, diverse content.

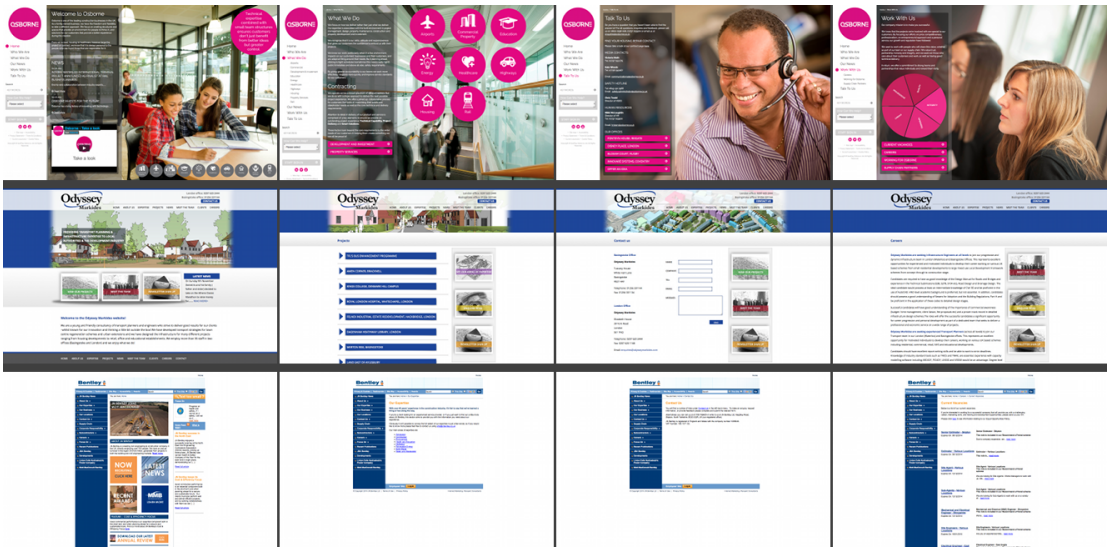


Figure 23. Examples of high (top row, osborne.co.uk, score = 1.39), medium (middle row, odysseymarkides.com, score = -. 18) and low (bottom row, jnbentley.co.uk, score = -.84) visual diversity. The scores are the averages of the six scaled diversity measures.

	Shopping apps		News apps		Business apps	
Diversity measure	r (df = 184 to 186)	p	r (df = 153 to 155)	p	r (df = 103 to 105)	p
CEDD	-.19	< .05	.30	< .001	.28	< .01
CL	-.19	< .05	.35	< .001	.16	.10
EH	-.01	.95	.23	< .01	.06	.57
SC	.11	.14	.50	< .001	.17	.08
FCTH	-.15	< .05	.29	< .001	.21	< .05
DC	-.18	< .05	.22	< .01	.20	< .05
Average	-.13	.07	.41	< .001	.24	< .05

Table 28. (Pearson's) Correlations between diversity scores and number of times an app was rated.

Contrary to our expectations, the visual diversity of apps in dataset 4 did not correlate with their ratings. An app rating (which is not the same as app popularity) combines a large number of factors, such as quality of service, usefulness, utility, functionality faults, or even, how much a new app GUI differs from the previous, already-learnt version. (A brief review of comments on Play Store reveals a large number of frustration reasons.) Many of these factors are clearly unrelated to app GUIs, which largely reduces the amount of rating variance that GUI-extracted metrics could explain. Datasets much larger than ours might be needed to discern a possible diversity-rating link.

Unlike app ratings per se, the number of app ratings and left comments might strongly depend on the initial, immediate aesthetics impression (Lindgaard et al., 2006; Miniukovich & De Angeli, 2014b). If favorable, such impression makes an app to stand out (Tractinsky, 2013) and increases the odds the app is installed (cf. the impact of aesthetics is the greatest during the initial use phases, Karapanos et al., 2009; Sonderegger et al., 2012). The user may later get disappointed with the app, and leave a low rating and negative feedback. However, the decrease is reflected in the app rating, and not in the number of ratings and comments. We, therefore, might assume app aesthetics to be carried over to the number of ratings (and possibly, the number of impulsive installs) to a much larger extent than to the app rating. Since visual diversity correlated well with aesthetics (cf., study

1, 2), one might reasonably infer it should correlate much stronger with the number of ratings than with ratings per se.

App popularity on Google Play – favorable attitude and much attention – depends on several factors, including app performance and usefulness, privacy and ethics policy, compatibility with other apps, quality of UX, users' comments, reviews and ratings in Google Play, and other aspects (Khalid et al., 2014). App popularity is also likely to depend on GUI quality and, as study 3 showed, on GUI visual diversity. We used imperfect descriptors of app popularity (the number of ratings depends on an app release date; both the number of ratings and average rating can fluctuate for different versions of the same app). Given the imperfections, the link that study 3 showed between screenshot-extracted scores and app popularity should be taken optimistically, but cautiously. The results of study 3 are the basis for future studies.

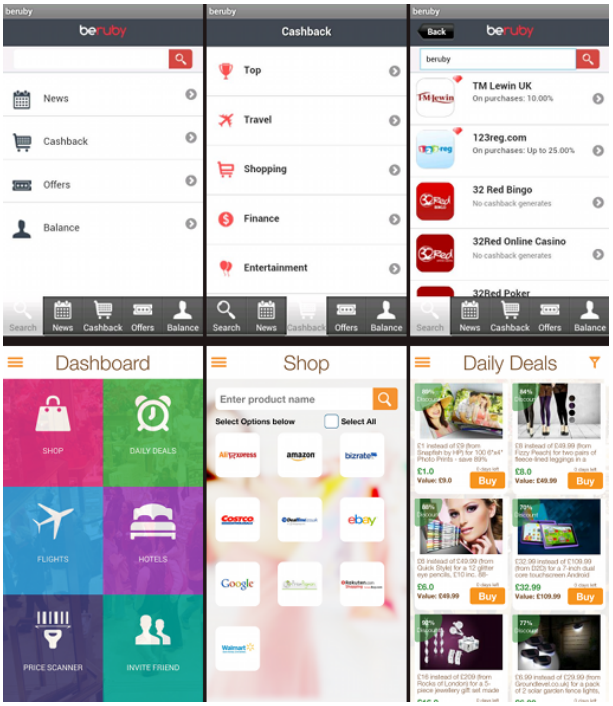


Figure 24. Examples of the least (top row) and most (bottom row) visually diverse shopping apps.

3.2.4 CONCLUSIONS

We have presented an initial exploration of GUI visual diversity in three studies. The studies have shown visual diversity to correlate with GUI visual aesthetics, and higher-level approach-avoidance tendencies (e.g., company favorability, study 2; the number of app comments, study 3). Given that these three studies used three different dataset, we might suggest that the link of diversity to GUI quality dimensions has not been a chance phenomenon. We believe the diversity-quality link, which is first studied on GUIs in this paper, is worth presenting to a wider audience.

Visual diversity is yet to be positioned within existing GUI quality frameworks (e.g., Hassenzahl, 2005; Rafaeli & Vilnai-Yavetz, 2004; Hartmann et al., 2008). However, already in this paper, we observed visual diversity to correlate with website aesthetics, and not with usability and content quality (study 2). We further observed the measures of visual diversity to perform well in both website (study 1 and 2) and mobile contexts (study 3). These initial observations may serve as the basis for the future studies on visual diversity and are one of the paper contributions.

We measured the diversity of websites and mobile apps using the well-tested algorithms from the image retrieval research field. In all three studies, the scores of individual measures converged, which suggested that all measures described the same overarching construct. Testing and applying the

algorithms in a novel context – not for image search and retrieval, but for measuring GUI visual diversity – is another contribution of this paper.

Figure 25. Examples of the least (top row) and most (bottom row) visually diverse news apps.

The present work should be extended in several ways. First, despite we observed a link between visual diversity and visual aesthetics in all three studies, we cannot make causality claims. A study needs to manipulate the level of diversity directly. Second, additional aspects of GUI quality should be included in a study of visual diversity, e.g., GUI visual complexity (Tuch et al., 2012) and online credibility (Beldad et al., 2010). Lastly, visual diversity needs to be studied in various use contexts. User quality criteria do depend on circumstances (Hartmann et al., 2008). The magnitude, or even, directionality of visual diversity impact may also depend on the circumstances. We might have observed such circumstance impact in study 3, Table 28. Further studies are needed.

3.3 Pick me! Getting Noticed on Google Play

Almost any search on Google Play returns numerous app suggestions. The user quickly skims through the list and picks a few apps for a closer look. The vast majority of the apps – regardless of how well-made they are – go unnoticed. App icons uniquely represent each app in Google Play and help apps to get noticed, as we demonstrate in the paper. We reviewed the visual qualities of icons that could make them noticeable and likable. We then computationally measured two of the qualities – visual saliency and complexity – for 930 icons and linked the computed scores to app popularity (the number of app ratings and installs). The measures explained 38% of variance in the number of ratings, if app genre was accounted for. Not only does such result assert the link between icon properties and app popularity, it also highlights the *automatic* prediction of app popularity as a promising research direction. HCI researchers, app creators and Google Play (or another mobile marketplace) will benefit from the paper insights on what antecedes app success and how to measure the antecedents.

3.3.1 INTRODUCTION

Millions of apps populate Google Play[11]. With the competition running sky-high, developers strive to convince the user to choose their apps: they polish the user experience (UX), promptly fix errors and add new functionality, and patiently respond to user complaints. However, all these aspects – which describe service quality, functionality and usability, but not look & feel –matter after the user has noticed, chosen and installed an app. First impression – which is almost entirely based on look & feel – is what makes the app to stand out, get noticed and get installed; first impression is what drives the initial success in all highly competitive IT markets (Tractinsky, 1997). Many apps fail to impress.

This paper offers a new line of research. Instead of taking user complaints (Khalid et al., 2014), menu structure (Chae & Kim, 2004) or app layout screenshots (Miniukovich & De Angeli, 2015b) as research input, we studied icons. Mobile app marketplaces (e.g., Google Play) rely heavily on icons to introduce apps to users. Sometimes more screen space is reserved for the icons than for titles, ratings and descriptions, in both desktop (Figure 26) and mobile versions (Figure 27) of Google Play. Icons have become *"the visual expression of a brand's products, services, and tools"*[12] and may well impact choose-and-install decisions.

We first reviewed the quality parameters of icons and selected two computationally quantifiable parameters: visual complexity and saliency. We then reviewed the automatic measures of visual complexity and saliency for images, graphical user interfaces (GUIs), and icons. Some of the measures could not be directly applied to icons. We customized them. Next, we scraped the popularity data and product icons of 943 Android apps from Google Play, and computed measure scores and matched them against app popularity scores. App popularity (as measured by the number of ratings) correlated with all complexity and saliency measures, with *contour congestion* and *amount of detail* being the two strongest predictors of popularity. With app genre as an independent variable, our linear regression models accounted for 38% of app popularity. Popularity indeed appeared to correspond to how much the icons caught attention. The results on app ratings (not the number of ratings) were much weaker, which was expected: the ratings reflect the overall post-use impression and depend little on the attention-grabbing qualities of icons.

In the remainder of the paper, we review related work on app success, icon quality, and automatic measures of visual complexity and saliency. We then list the automatic measures selected for the study, describe the study and discuss results. The implications for the practitioner and future research are offered in the end.

[11] https://play.google.com/store/apps

[12] http://www.google.com/design/spec/style/icons.html

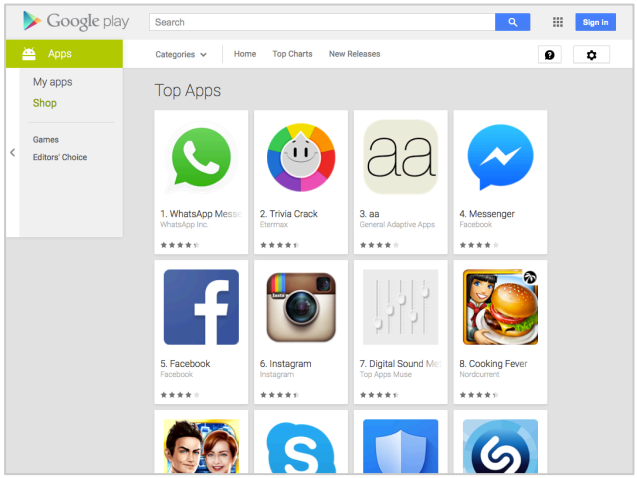


Figure 26. A listing of apps on Google Play as seen from a wide-screen device.

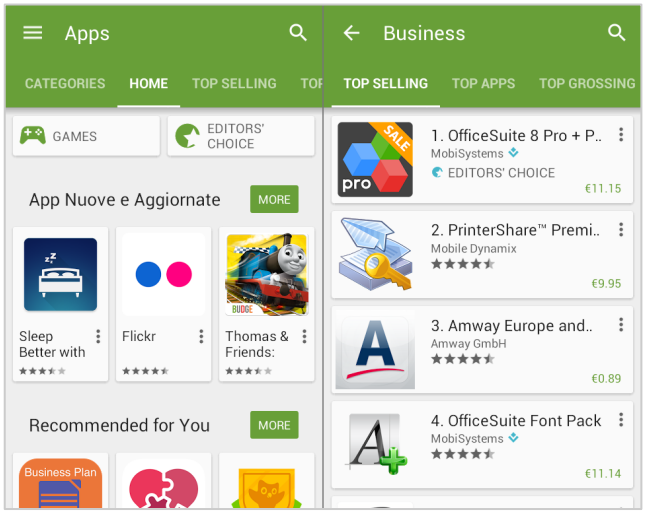


Figure 27. App listings on Google Play as seen from a mobile device.

3.3.2 Related work

Within the research on UX, mobile apps stay a less explored domain, with the majority of research done on websites. However, mobile apps offer a unique research opportunity. Unlike websites, apps reside within large virtual marketplaces (Google Play being one of them), which host apps, let users evaluate them, and track app search, usage and preference statistics. Unlike for websites, the statistics for apps are tracked in exactly the same manner and under the same conditions: users search apps on the same website, rate apps using the same rating mechanisms, and receive app suggestions from the same suggestion algorithms. Such homogeneity makes collected data ideal for analyses. A few researchers did leverage the marketplace advantages; they showed a profound effect of user feedback (app reviews and ratings) on app success (Mudambi & Schuff, 2010; Kim et al., 2011), and reviewed the types of user complaints showing their effect on app success (Khalid et al., 2014). Other researchers looked at individual apps, outside of marketplaces. They delved in app user behavior and found stable behavioral patterns (Böhmer et al., 2011), linked the structure of navigation menus and user satisfaction (Chae & Kim, 2004), and tried to computationally model app GUI appreciation using app screenshots (Miniukovich & De Angeli, 2015b). Many of these efforts dealt with app quality, but few directly addressed what makes the user like and choose a visual item (e.g., an app) from a list – favorable first impression.

3.3.2.1 First Impression

The models of user experience evolution distinguished three stages of UX (Sonderegger et al., 2012): orientation (the user chooses an app and explores it), incorporation (the user incorporates the app in her daily activities), and identification (the user forms emotional attachment to the app). At all stages, the user appreciates beauty in the visual appearance of their phones and apps (Ling et al., 2007). However, the appreciation particularly prevails during the first stage, orientation, when the user chooses to install and stick with an app (Sonderegger et al., 2012).

The appreciation of visual aesthetics happens very quickly, in under a half second (Lindgaard et al., 2006; Miniukovich & De Angeli, 2014a), changes little after it is formed (Tractinsky et al., 2006), and affects user actions (Thielsch et al., 2013). In a study on websites, Thielsch et al. (2013) linked first impression to the overall impression, intention to revisit and intention to recommend to friends. Kim and Fesenmaier (2008) linked first impression to the stay/leave decision. They described a typical action flow of looking for a website: the user queries a search engine and gets a list of results, opens a link from the list, has a brief – a few seconds long – look at the website, and decides if she stays on it or goes to another website down the list. The action flow of looking for apps would be similar to that for websites, and the decision which app to choose would also depend on a positive first impression. Two qualities that form such impression are high saliency (McCay-Peet et al., 2012) and low complexity (Reber et al., 2004).

3.3.2.2 Visual Saliency and Complexity

Among the qualities that constitute a good icon (e.g., McDougal et al. (2000) listed three such qualities: icon concreteness, visual complexity and distinctiveness) two qualities – saliency and complexity – are vision-based and can be relatively easily estimated. Icon saliency corresponds to the icon capacity to stand out from a row of icons. Saliency cannot be estimated in isolation; it depends on the other icons in the surroundings of target icons. For example, higher luminance contrast between an icon and its background would make the icon to stand out, i.e., would increase its saliency (McDougall et al., 2000). Other visual features that also could contrast an icon from its surroundings include color, edge orientation, texture, size, motion and flicker (Wolfe & Horowitz, 2004). Besides making an icon visible, higher saliency could result in a more positive affect and favorable impression, particularly, if the user had already been looking for the icon (McCay-Peet et al., 2012).

The other visual feature of icons, their visual complexity, was investigated more often than any other feature (Ng & Chan, 2008) and was even attempted to quantify automatically (Forsythe et al., 2003; Forsythe, 2009). Icon complexity corresponds to the amount and intricacy of detail within an icon (cf., Ng & Chan, 2008), and has been linked to a range of outcomes. For example, McDougal & Reppa (2013) linked icon visual complexity to visual appeal and search time. They explained the link via the concept of processing fluency (Reber et al., 2004): lower complexity led to lower effort from the user, which then led to liking and higher appeal ratings. Such subconscious attribution of simplicity to beauty happens almost instantly. Using websites as stimuli, Tuch et al. (2012) demonstrated the attribution on a range of exposure intervals, from 17ms to 1000ms; the impact of complexity was evident already after the 17ms exposures.

A few researchers took a descriptive approach to complexity and listed a number of complexity dimensions. Oliva et al. (2004) looked at the verbal descriptions of indoor photographs and listed six dimensions: amount of detail, objects and colors, visual clutter, symmetry, open space, organization and contrast. Miniukovich & De Angeli (2014a) studied the immediate impressions of webpage complexity. They listed eight complexity dimensions – color variability, clutter, contour congestion, contrast, symmetry, grid quality, ease of grouping, and prototypicality – and suggested automatic measures for all dimensions but the last three. A definitive taxonomy of complexity is still to be proposed.

3.3.2.3 Metrics and Measures

Both *metric* and *measure* refer to estimates and estimating. However, we follow the tradition (Black et al., 2008) of calling direct automatic measurements a measure and higher-level combinations of measures a metric. Metrics are easier to interpret, more reliable than measures, and thus, more desirable.

Visual saliency has been well discussed in the literature on modeling human visual perception, which resulted in several computational models. (Among them, Itti et al.'s, 1998, model of saliency-

based visual attention – and ensuing metric – is one of the most used.) The majority of models and measures of visual saliency rely on target-surround comparisons on one or several dimensions (cf., Parkhurst & Niebur, 2004; Itti et al., 1998). The measures first take an image as an input and scale it several times (e.g., using the dyadic Gaussian pyramids). The scaled versions of the image are then compared against each other: each pixel of finer-scale version (considered a center) is compared against the corresponding pixel of coarser-scale version (considered a surround). Bigger difference between the two pixels corresponds to higher saliency in that area. Any visual feature can be used for the pixel comparison, with pixel color, luminance, and edge orientation being most popular.

Visual complexity has attracted much attention in HCI, which resulted in a multitude of measures for GUIs, or images, photographs, and icons. The measures for GUIs often include knowing the specificity of GUIs (e.g., buttons are rectangular and placed parallel to the screen sides) and make sense only in the context of GUIs (natural images rarely have rectangular carefully arranged objects). Harper et al., (2013) looked at the top-left corners of webpage elements: more corners and less uniform distribution of corners on a page were associated with higher complexity. Nadkarni & Gupta (2007) also looked at the underlying structure of webpages to estimate complexity. Among many measures, they counted webpage graphics, words, colors, links and pop-up ads, and computed the average download time, percentage of white space and website structure depth. Other measures rely on the analysis of rectangular structure of GUIs. Thus, Wu et al. (2013) analyzed complexity by slicing webpages in rectangular blocks and looking at their number, width, height, width-to-height ratio, average colorfulness and brightness, texture and other features. Miniukovich & De Angeli (2015a) sliced webpage and mobile app screenshots in rectangular GUI blocks and used the blocks to estimate grid quality and symmetry. Higher quality and symmetry were found to correspond to higher aesthetics ratings.

The complexity measures can be applied alike to artistic photographs, screenshots of GUIs or icons, since they impose no restriction on their input. Purchase et al. (2012) tested several such measures on the photographs of objects and found measure scores to correlate with the user scores of complexity. The measures included the number of colors before and after color reduction, number of contour pixels, the variance of pixel luminance, and sizes of image files in the JPEG, PNG and GIF formats. The file size of compressed images is often used as the simplest, most easily available complexity measure (e.g., in Tuch et al., 2012; Wu et al., 2013; Forsythe, 2009; Harper et al., 2013). The image compression algorithms remove redundancy from the original images (e.g., by encoding large same-color areas with only few bytes); more redundancy to be removed corresponds to lower complexity and lower compressed image size. Several research teams (Reinecke et al., 2013; Zheng et al., 2009; Forsythe et al., 2003) used a quadtree decomposition to estimate the complexity of webpages or computer icons. Such decomposition keeps splitting an image in blocks till the pixels within a block are homogeneous enough (e.g., they all are of approximately the same color). More blocks correspond to higher complexity. In addition to quadtree decomposition, Forsythe et al. (Forsythe et al., 2003) counted the number of perimeter and edge pixels of icons. All three measures – the number of quadtree blocks, perimeter pixels, and edge pixels – were strongly intercorrelated, and correlated with the user scores of icon complexity. Finally, Miniukovich & De Angeli (2014a) estimated a number of dimensions of webpage complexity. They looked at webpage color variability, visual clutter, contour congestion, figure-ground contrast, and symmetry. All measure scores correlated with the user complexity scores.

3.3.3 Study preparation

We expected app success to partially depend on icon quality. This idea was split in three hypotheses and tested in a study. We first sampled app icons and app popularity data from Google Play. We then adapted several methods of saliency and complexity computation for the use on icons. Lastly, we computed saliency and complexity scores and matched them against the popularity data, which validated the methods and tested the hypotheses.

3.3.3.1 Hypotheses

Past work suggested that mental attention-guidance mechanisms might largely reduce the input that humans process consciously (Wolfe & Horowitz, 2004; Mormann et al., 2012). The input might then be further reduced down to what is considered simple, and therefore, likable (Lindgaard et al., 2006; Reber et al., 2004). We applied these ideas in a schema of app selection, Figure 28. When the user

freely browses an app listing, icon saliency and complexity reduce the initial abundance of apps down to few. The few selected apps may then be launched in the loop, where having more users generated more attention, which generated more users. We equaled the number of users to app popularity and formulated the first hypothesis, *"Icon saliency and complexity are related to the number of app users"*.

The schema (Figure 28) did not presume – though did not exclude – the link between icon visual features and overall app appreciation (as measured by the mean app rating). If existent, we would expect such link to be weak, since many other app quality factors (e.g., app utility and usability, quality of in-app ads, or developer responses to users' requests) would largely dilute the initial, vision-based icon impression. We formulated the second hypothesis as *"Icon saliency and complexity are unrelated or weakly related to users' ratings of apps"*.

A study of app popularity should consider app genre. Different genres target different crowds and simple counting of users may be an unsuitable metric of popularity. For example, travel apps may be used by few travelers and uninstalled quickly after the travel, whereas media apps (e.g., a music player) are used by everyone all year round. Both apps may be popular but the number of users is very different. The same should not be true for app appreciation: apps can be appreciated (i.e., have a high rating from users) regardless of their genre. We formulated the final hypothesis as *"App genre is related to the number of app users, but unrelated to users' ratings of apps"*.

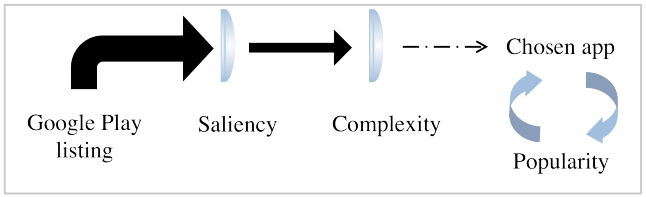


Figure 28. The schema of app selection: in the absence of top-down selection factors (e.g., task at hand, (Wolfe & Horowitz, 2004; Navalpakkam & Itti, 2005)), the bottom-up, vision-based factors determine the choice.

3.3.3.2 Stimuli Sampling

We sampled production icons of apps from Google Play. App users can browse Google Play and choose apps from tile-like app collections (Figure 26, Figure 27). The collections are ordered according to either popularity (app rating, number of installs, ratings and comments, and possibly, several more criteria) or the match between a search query and app description. We only sampled apps from the latter, search-match based collections. The former, top-app collections were avoided because Google did not disclose their principles of assembling those collections – any systematic biases (e.g., small apps or unprofessionally-looking apps could not appear on those lists) would not be canceled out, even by our random app selection and large samples.

We selected 943 apps in six categories: shopping (221 apps), education (102 apps), business (121 apps), news and magazines (158 apps), travel and local (182 apps), and media and video (159 apps). The search query for each category comprised of the name of category followed by the word *app* (e.g., *shopping apps*). We developed an extension for Mozilla Firefox, which automatically entered the search queries and collected data from Google Play. The extension operated within a clean, newly created Firefox profile, which ensured Google Play could not customize search results based on authors' search history, browsing history, preferred language or any other personal data.

For each app, we collected its product icon, rating, number of times the app was rated, and number of installs. Product icons were 170×170 pixel images in the PNG format, 32 bit per pixel; app ratings were decimal numbers, from 1.0 to 5.0; counts of ratings were integers; counts of installs were ranks (e.g., *'500 to 10,000'* had a higher rank than *'100 to 500'*). The apps with ratings below 2.0, 13 in total, were removed from the sample. The sampling happened in February 2015.

3.3.3.3 Design

The study included three dependent variables (mean app rating, number of ratings, and number of app installs) and ten independent variables (app category, icon saliency, number of dominant colors, number of contour pixels, contour congestion, contrast, symmetry, number of quadtree blocks, number of high-pass contour pixels, and contour energy).

3.3.4 Computation of Saliency and Complexity

This section reports our computational methods. The research on visual saliency – but not on visual complexity – has offered higher-level metrics. We considered and used one saliency metric and a selection of visual complexity measures.

3.3.4.1 Saliency

We relied[13] on Itti et al.'s model of saliency-based visual attention (Itti et al., 1998). The model mimics human perception; it takes an image as an input, computes target-surround differences for three visual features – color, luminosity and edge orientation – and combines the differences across the features. The final output is a saliency map; brighter areas correspond to more abrupt target-surround differences, i.e., to higher saliency. Since the model analyzes several visual features (namely, color, luminosity and orientation), we call its output a metric, not a measure.

Estimating saliency of an icon required placing the icon into meaningful surroundings resembling the layout of Google Play. In Google Play (Figure 26, Figure 27), icons differentiate one app "tile" from another; the look of other features – such as labels, stars and white tiles – is constant for all apps. We decided to skip the other features and generated target-surround images from icons only, ten images per icon (Figure 29). Knowing which icons co-occurred with a target icon was not possible and we drew surround icons randomly. The target icon was placed in the center of white canvas, 24 other randomly-selected icons were placed around the icon, ten pixels of white space was kept constant between the icons. We then computed saliency maps (70 × 70 pixel, three visual features – color, luminance and orientation – were used, the global center bias was not modeled) for the ten target-surround images, sliced the maps back into 25 squares and summed up the values within each square. The sums were ranked in descending order. The rank of the target icon (Figure 29, in the red square) was averaged across ten target-surround images and taken as the metric of saliency.

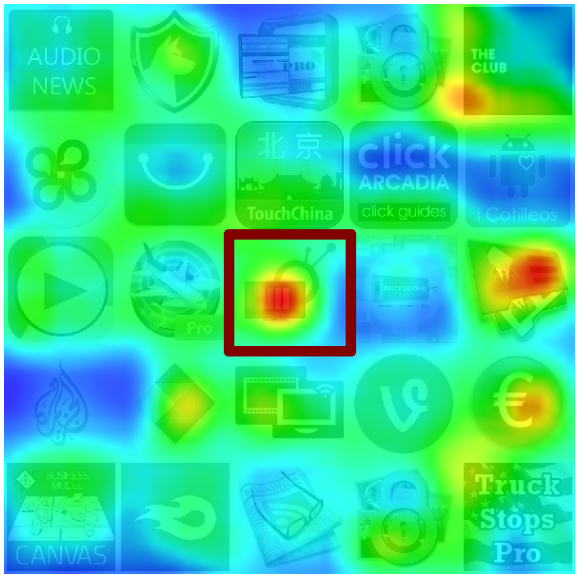


Figure 29. The target icon (in a red square) surrounded by 24 other icons served as an input for the saliency metric. The corresponding saliency map is overlaid as a heatmap.

3.3.4.2 Complexity

Before the study, we considered many of the known measures, but selected and tested only one measure per complexity dimension (namely, color variability, amount of detail, congestion, contrast, symmetry) and several extra measures, which have been applied specifically to icons in the past

[13] We re-used Jonathan Harel's implementation of the saliency algorithm (Itti et al., 1998), which he kindly published online, http://www.vision.caltech.edu/~harel/share/gbvs.php

(quadtree decomposition and high-pass filtering). We included the extra measures to link up our work with past work (Forsythe, 2009; Forsythe et al., 2003; Thielsch & Hirschfeld, 2010; 2012).

3.3.4.2.1 Color Variability

Color variability, consists of two aspects: number of dominant colors and color range (Miniukovich & De Angeli, 2014a). Dominant colors occupy a significant portion of image; a human can easily count them with a naked eye. Color range describes the whole multitude of color shades and tones, which may go unnoticed by a human. The number of dominant colors had been shown to negatively correlate with webpage complexity scores (Miniukovich & De Angeli, 2014a) and we re-applied the idea to icons We counted the number of dominant colors using the method of uniform color quantization (cf. (Miniukovich & De Angeli, 2014a; 2014b; 2015a)): all pixel color values (in RGB) of an icon were put in a color cube; the cube was then sliced in 512 sub-cubes; all sub-cubes that contained at least three values were counted; the counts were taken as dominant color estimates. Figure 30 demonstrates the color quantization: the main colors are shown under their two corresponding icons; the size of color patch corresponds to the proportion of icon that the color occupies.

Figure 30. Color quantization. Color semi-tones and shades from two icons were discarded and only the main colors were counted.

3.3.4.2.2 Amount of Detail

Researchers have suggested several measures of amount of detail, such as Feature Congestion (an estimate of used feature space, e.g., color or luminance space, (Rosenholtz et al., 2007)), Subband Entropy (an estimate of feature redundancy in an image, (Rosenholtz et al., 2007)), image file sizes in JPEG (Tuch et al., 2012), and number of contour pixels (Rosenholtz et al., 2007). The measures were observed to strongly intercorrelate (Miniukovich & De Angeli, 2014a; 2014b). We chose to compute the number of contour pixels: it was the simplest measure, which was also tested on icons (Forsythe et al., 2003). Contours were detected using the Canny edge detector (low threshold - .11, high threshold - .27; cf., (Rosenholtz et al., 2007)). The original icons were converted into grayscale icons as contour pixels were detected and counted; the counts were normalized by icon sizes and taken as the estimates of amount of detail. Figure 31 shows the contours of an icon.

3.3.4.2.3 Contrast

Luminance contrast describes the difference in luminance between two adjacent image areas. Lower contrast increases the effort to make sense of image, and thus, increases its complexity. As an example, Hall & Hanna (2004) demonstrated higher text-background contrast to decrease the effort of reading. We used the measure of contrast from (Miniukovich & De Angeli, 2014a; 2014b). First, icon edges were detected at several consecutive thresholds using the Canny edge detector. The thresholds we used for icons (from .25 to .95 with the step of .1; the low threshold was always 40% of the high threshold) were higher than the thresholds proposed for websites (from .10 to .70, (Miniukovich & De Angeli, 2014a)) because, we observed, icons often contained only high-contrast edges. Such edges would always be detected unless higher thresholds were used. Subtler edges (i.e., the edges detected at lower thresholds) were assigned higher weights; stronger edges were assigned lower weights. (The weights ranged from 0 to 1.) Weighted edge pixels were counted and normalized by the entire number of edge pixels. The normalized count was taken as an estimate of

icon contrast. Figure 32 demonstrates the contrast measure: stronger edges are green, subtler edges are red.

Figure 31. Icon contour detection. Contour pixels are counted and taken as an estimate of amount of detail.

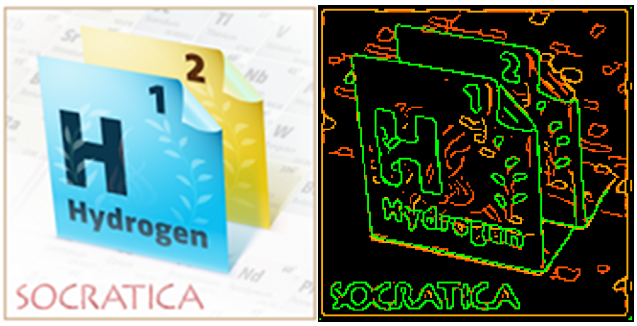


Figure 32. Contrast measurement. The transition from green to red reflects lowering contrast.

3.3.4.2.4 Symmetry

Vertical mirror symmetry facilitates shape perception (Machilsen et al., 2009) and has been shown to increase the appeal of patterns (black-white dots, lines and shapes, (Tinio & Leder, 2009)) and webpages. For GUIs, Researchers developed several automatic block-based measures of symmetry (Balinsky, 2006; Miniukovich & De Angeli, 2015a). These measures could not be applied to icons: they required sets of rectangular blocks (e.g., GUI elements, such as texts or buttons) as inputs, whereas icons rarely consisted of rectangular blocks. We instead turned to a method[14] of detecting local symmetries from the object detection and recognition domain (Loy & Eklundh, 2006). The method generates a set of descriptors (based on the SIFT keypoints), tries to pair the descriptors, and returns a set of symmetric pairs and associated symmetry axes. The axes could be of any position and tilt, not only the central vertical axis. The method, however, performed unsatisfactorily. Our icons were too simple images, which resulted in few to none SIFT keypoints detected per icon. In some cases, the low number of keypoints did not allow symmetry estimation at all. In other cases, a visual inspection of icons and their symmetries suggested a link between the number of keypoints and amount of detail in an icon, which was confirmed. The estimates of the amount of detail – measured as the count of contour pixels (cf., (Rosenholtz et al., 2007)) – correlated with icon symmetry estimates by the method (Loy & Eklundh, 2006), $r(497) = .52$, $p < .001$. Symmetry should have been independent of the amount of detail.

We finally turned to a contour-based measure of global vertical symmetry, which was first developed for GUIs (Miniukovich & De Angeli, 2014a). The measure detects contour pixels (the Canny edge detector, low threshold = .11, high threshold = .27) and uses them as keypoints. The keypoints were marked as symmetrical if they had a matching keypoint across the central vertical axis in a 2-pixel radius area. (Miniukovich & De Angeli, 2014a, used 4-pixel radius areas, which we

[14] We used Loy et al.'s (Loy & Eklundh, 2006) implementation of the method, from http://www.nada.kth.se/~gareth/homepage/local_site/code.htm

reduced to reflect a small icon size.) The ratio of symmetrical keypoints to all keypoints was taken as an estimate of icon symmetry. Figure 33 demonstrates the symmetry measure. Symmetrical keypoints are green; asymmetrical keypoints are red.

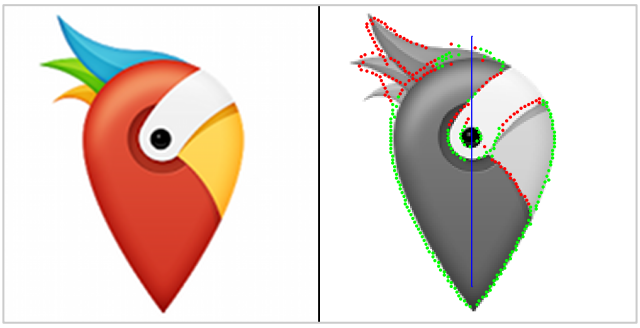


Figure 33. Symmetry measure. Symmetrical contour points (green) have a pair across the central vertical axis (the blue line). Non-symmetrical points (red) have no pair.

3.3.4.2.5 Quadtree decomposition

Quadtree decomposition describes the homogeneity of images. An image is iteratively split in square blocks till a feature of block pixels (e.g., luminance or color) varies within the block by less then a threshold. We used luminance as the pixel feature, 25% of maximal luminance as the threshold, and 4 pixels as the minimal block size (Figure 34). The size of images for quadtree decomposition needs to be a power of 2: we upscaled our icons to 256×256 pixel sizes using the bicubic interpolation. Quadtree decomposition produces many small blocks at around image contours (see Figure 34); the number of quadtree blocks strongly correlates with other measures of image detail (Forsythe et al., 2003), e.g., with the counts of contour pixels that we already considered above. However, we still included the number of quadtree blocks in the study as a link to the past work (Forsythe et al., 2003; Zheng et al., 2009; Reinecke et al., 2013; Wu et al., 2013).

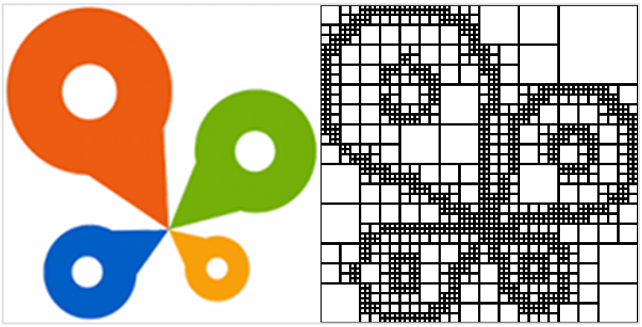


Figure 34. Quadtree decomposition. An image is split in square blocks till the pixel luminance within a block varies less than 25% of maximal luminance.

3.3.4.2.6 Spatial Frequencies

High spatial frequencies describe the fine detail within an image. Researchers argued that humans might have evolved to effortlessly perceive the absolute luminance levels (i.e., low frequencies), but need a significant effort perceiving small detail (i.e., high frequencies). Past research indeed linked high spatial frequencies to image complexity (Forsythe, 2009) and GUI aesthetics (Thielsch & Hirschfeld, 2012; 2010). However, past research had only used high-pass filtered images, and not quantified high-frequency information automatically. We filtered our icons with a Gaussian low-pass filter (kernel size = 5, sigma = 1) and subtracted the filtered versions (Figure 35, center) from the original icons (Figure 35, left) to get the high-frequency information (Figure 35, right). We then considered two high-pass based measures: the number of non-zero pixels (all dark pixels, Figure 35 right) and the average luminance of non-zero pixels (cf., Yu et al.'s (2013) measure of edge energy: a similar filtering idea, but based on the Sobel kernels). The latter measure describes the sharpness of contour-background difference (Figure 36); we titled it as contour energy.

Figure 35. Frequency filtering. The central image is processed with a low-pass filter; the right image - with a high-pass filter.

3.3.4.2.7 Congestion

Contour congestion – too many contours too close or overlapping each other – requires an observer to focus her fovea vision on each image patch; she cannot grasp the image meaning using only her peripheral vision (van den Berg et al., 2009). We used the contour congestion measure from (Miniukovich & De Angeli, 2014a). First, for each of three RGB channels, we found pixel pairs with the value difference of more than 50 and marked then as edge pixels. One-pixel thick contours were marked twice. We then counted the edge pixels with at least two other contours in their 20-pixel proximity. The counts were normalized by the number of all edge pixels and taken as the measure of contour congestion. Figure 37 demonstrates the measure; congested contour pixels are red, non-congested contour pixels are green. A review of congestion score histogram showed a non-normal distribution of the scores: too many icons with little to no congestion. To counteract this, we randomly added one edge pixel per icon line and recomputed the measure. The recomputed scores correlated strongly with the original scores ($r(928) = .96$, $p < .001$), but the distribution of recomputed scores was close to normal – we used them in the further analysis. Finally, congestion scores strongly correlated (Pearson's r from .67 to .77, $p < .001$) with the contour pixel counts, number of quadtree blocks and number of high-frequency pixels – three measures that describe the amount of detail or "set size", a psychology concept to quantify the amount of information in a display (cf., (Rosenholtz et al., 2007)). We normalized the congestion scores by the number of quadtree blocks, and thus, decoupled them from the set size.

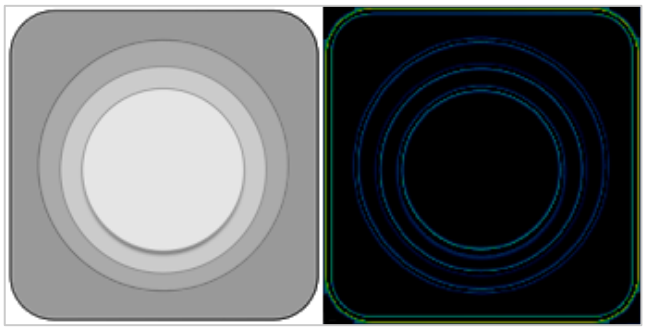


Figure 36. Contour energy. The brighter and greener contours of the processed icon (on the right) correspond to the larger visual contour-background difference of the original icon (on the left).

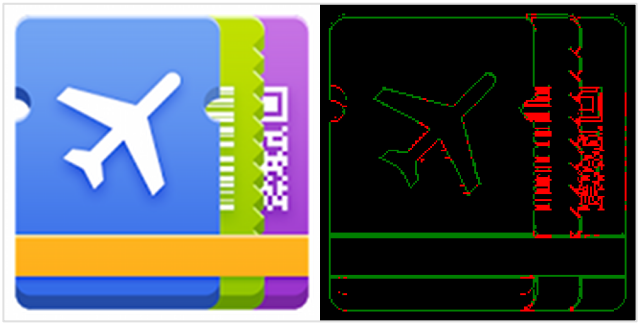


Figure 37. Contour congestion. Congested contours are marked red.

3.3.5 Study Results

We applied the automatic metric and measures to the collected icons. The resulting scores were then compared against app appreciation and popularity data, which let us test the hypotheses and validate the automatic measures.

3.3.5.1 Data Preparation

We first filtered out outliers from the dataset (values that deviated from the mean by more than 3×SD; only few such values were found, e.g., 8 out of 930 values for the measure of dominant colors). We then reviewed the histograms of all variables. One dependent variable – the number of app ratings – was strongly positively skewed (Figure 38). We log-normalized it because a 10^4-fold difference in rating counts (some apps had millions of ratings, while some others less than a hundred) obviously would not correspond to a 10^4-fold difference in icon quality and therefore Pearson's correlation would not describe well the connection between ratings and icons. The distribution of the other numerical variables approximated the normal distribution. Lastly, we reviewed the Lowess curves (they resemble smooth, curvy regression lines since they are re-computed for each local region). The curves revealed no non-linear dependencies between dependent and independent variables.

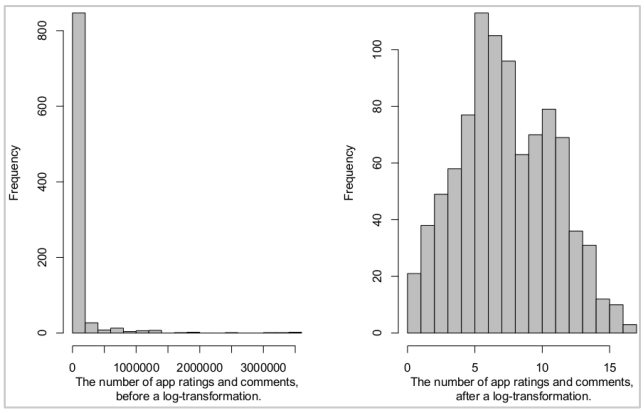


Figure 38. A review of the histogram of rating counts (left) shows the vast majority of apps had relatively few ratings; the data needed to be log-transformed (right).

3.3.5.2 Measure Validation

A review of cross-correlations among automatic measures showed that the count of contour pixels, number of quadtree blocks and number of high-frequency pixels tended to strongly cross-correlate (Pearson's r from .71 to .84, $p < .001$). From these three, we chose the number of high-pass contour pixels for further regression analysis, which reduced our set of computed independent variables from nine to seven. The cross-correlations among the seven variables were acceptable for the use in a regression model (Table 29).

	Congestion	Symmetry	Contrast	Saliency	High-pass contours	Contour energy
Dominant colors	.13***	.00	.50***	.04	.56***	-.24***
Congestion	--	-.17***	.17***	.23***	.14***	.20***
Symmetry	--	--	.00	-.04	-.07*	-.07*
Contrast	--	--	--	.24***	.56***	-.22***
Saliency	--	--	--	--	.26***	.05
High-pass contours	--	--	--	--	--	-.24***

* $p < .05$; *** $p < .001$.

Table 29. The cross-correlations among computed complexity measures and a saliency metric.

3.3.5.3 Hypothesis Testing

Two independent variables – the mean app rating and number of ratings – did not correlate ($r(928) = .04$; $p = .17$). The number of install (a categorical variable) strongly correlated with the number of ratings (Spearman's $r = .97$; $p < .001$). We omit reporting further results on the number of installs since such results would mirror the results on the number of ratings, but would also involve a more complicated statistical analysis (the number of installs was an ordinal variable, not a ratio variable). We instead report a simpler, conventional linear regression on the number of ratings.

The number of app ratings correlated with all automatic measures (Table 30); the mean app ratings correlated only with the measures of contrast ($r = -.08$, $p < .05$), amount of detail ($r = -.07$, $p < .05$) and congestion ($r = -.12$, $p < .001$). Such evidence supported hypotheses 1 and 2. The effect of app genre was significant on both dependent variables, but much stronger on the number of ratings ($F(5,924) = 76.96$, $p < .001$, $\square^2 = .29$) than on the mean app rating ($F(5,924) = 8.37$, $p < .001$, $\square^2 = .04$). This supported hypothesis 3. Seven computed independent variables were put in a stepwise regression with backward exclusion to select the best linear model (Table 31). The inclusion of app genre – a categorical, non-computed variable – in the linear model further improved model fit (R^2) up to .38, Table 31. All genres but one contributed to the fit significantly. Further detailing the between-genre differences falls outside the scope of paper, and we leave it out.

Computed variables	Pearson's r
Dominant colors	-.14***
Congestion	-.28***
Amount of detail	-.30***
Symmetry	.08*
Contrast	-.25***
Saliency	-.19***
High-pass contours	-.27***
Contour energy	-.17***
Number of quadtree blocks	-.24***

* $p < .05$; *** $p < .001$.

Table 30. Pearson's correlations between the number of app ratings and computed measures, df varies from 920 to 928.

Predictors	β	t
Congestion	-.17***	-5.40
Contrast	-.14***	-3.71
High-pass contours	-.21***	-5.58
Contour energy	-.22***	-6.67
Saliency	-.06[1]	-1.96
R^2 (R^2_{adj})	.19 (.18); $F(5,907) = 41.79$***	
With *app genre* as a predictor		
R^2 (R^2_{adj})	.38 (.38); $F(9,903) = 55.93$***	

*** $p < .001$; [1] $p = .05$

Table 31. A linear regression model of app popularity. (The outcome variable is the number of app ratings.)

3.3.6 Study Discussion

The study has linked the popularity of mobile apps (the number of installs and number of ratings) to the visual features of app icons (saliency and complexity) and tested a method of icon saliency and complexity computation.

3.3.6.1 App Popularity

We hypothesized that users may chose apps because of their visually salient but simple icons design (cf., Figure 28). To test such a hypothesis, we calculated eight visual-complexity measures and a single saliency metric, and matched them against app popularity data. When combined in a linear regression model, the measures, metric and app genre explained 38% of variance in the number of app ratings, Table 31. Such results supported hypothesis 1; they suggested that the visual properties of icons might indeed be linked to app popularity.

Following hypothesis 2, we did not expect the same correspondence for the mean app rating – another marker of app quality. The computed scores only weakly correlated with the mean app ratings (congestion correlated the strongest, $r = -.12$, $p < .001$) and at best accounted for 1% of rating variance. Such results supported hypothesis 2 and did not surprise. The user might choose and install an app, but then - despite the app icon was nice and catchy – dislike it for a host of reasons: usefulness, look and feel, marketing campaigns, update frequency, communication between developers and users, GUI usability and aesthetics, and many other factors (Khalid et al., 2014).

Hypothesis 3 has also appeared to be supported: app genre explained much of app popularity (the size of effect on the number of ratings was $\square^2 = .29$), and little of app appreciation (the size of effect on the mean ratings was $\square^2 = .04$).

3.3.6.2 Computational Method

To estimate image visual saliency, we extended the well-known method from Itti et al. (1998). The method assumes the presence of both target and surroundings, i.e., saliency cannot be estimated in isolation; meaningful surroundings are needed. We created such surroundings by placing each icon (a target) on a white canvas and wrapping the icon with other, randomly selected icons (surroundings). The canvas was then fed in the algorithm (Itti et al., 1998) to compute a saliency map. The saliency of target icon was then carved out from the map. Such computation was repeated ten times and resulting saliency values averaged, which should have reduced random error.

Our extension of Itti et al.'s (1998) method let us calculate icon saliency relative to other icons (i.e., to account for the meaningful surroundings). Such approach appeared fruitful as the resulting saliency values correlated with the number of app ratings ($r = -.19$, $p < .001$). The direction of correlation was as expected: higher saliency rank (lower saliency) corresponded to fewer app ratings. The effect of saliency on popularity stayed significant though diminished ($p = .05$) after accounting for visual complexity (Table 31). We might speculate visual saliency indeed selected candidate-objects for further, complexity-based mental processing (Figure 28, cf., Wolfe & Horowitz, 2004).

We estimated icon visual complexity with eight measures; all measure scores correlated with the number of app ratings (Table 30). Three measures – the number of contour pixels, number of quadtree blocks, and number of high- pass contour pixels – described the same concept, known as *set size* in psychology (cf., Rosenholtz et al., 2007). As expected (cf., Forsythe et al., 2003), the measures strongly cross-correlated and were included in the analysis as a link to the past work (namely, Forsythe et al., 2003; Forsythe, 2009). All three *set-size* measures correlated negatively with the number of app ratings. The measures of contour congestion, contrast, contour energy, and dominant colors also correlated negatively the numbers of app ratings (Table 30). Only the measure of symmetry correlated positively with those numbers. Symmetry, however, was a weak predictor of app popularity ($r = .08$, $p < .05$), implying either the need for a symmetry measure better than ours or the low importance of symmetry for the user. The latter would corroborate the results on webpage aesthetics (Tuch et al., 2010; Miniukovich & De Angeli, 2014a, the impact of symmetry was weak or conditional). Lastly, the number of dominant colors correlated relatively strongly with the three measures of set size (e.g., with the number of high-frequency pixels, $r = .56$, $p < .001$). This might follow from the icon specificity, when each new element in an icon tended to have a unique color.

3.3.6.3 Implications

We believe our findings and algorithms could be applied in several domains, e.g., as an insight source about the potential of apps to succeed (investment decisions) or as a part of the Google Play procedure for selecting top-quality apps. (Such procedure is a part of app search and helps the user find best apps.) HCI researchers could substitute user data collection with our measure computation, and thus, speed up their research on icons and mobile apps. Logo and icon designers could draw informed design insights from the paper findings. The multitude of requirements to satisfy in an icon design (e.g., linking the icon to company name or mission; being original; staying within a limited screen space; or complying with the general visual style of app) might carry the designers away from creating icons that attract users. We might suggest to the designers to make icons noticeable (e.g., use less common color combinations and line directions; also see (Kumar et al., 2013) on what is common on the Web) and simple (e.g., use little of intricate, fine-grained detail; spread the detail across the icon; use fewer main colors; use semitones and shades of main colors, and antialiasing to

create "softer" lines). However, converting our findings in a more universal set of guidelines or design-evaluation tools requires further work.

The present work can be extended further. We showed app icon complexity and saliency to play a role in app selection decisions. Other visual features of icons could play a similar role. For example, specific colors (cf., Reinecke et al., 2013) or textures (Tamura et al., 1978) might be preferred in some cultures and carry over the preference on to app selection; or arts-based regularities (the rule of thirds, use of complementary colors, or golden ratio) might convey the aesthetics of icons. Cognitive features, such as familiarity with a brand or brand value, could contribute to app popularity and should also be explored (De Angeli et al., 2008). Finally, a future study may computationally address visual appeal – a quality shown to make an icon to stand out, particularly in stressful situations (Reppa & McDougall, 2015).

3.3.7 Conclusion

The paper offers two main contributions. First, we have demonstrated a link between the visual features of mobile app icons – namely, visual complexity and saliency – and mobile app popularity. To the best of our knowledge, this paper is the first such demonstration; past efforts concentrated on, for example, user complaints (Khalid et al., 2014) or in-app visual consistency (Miniukovich & De Angeli, 2015b). Second, we assembled a set of computational methods, which could estimate icon complexity and saliency. The set included measures from the research on the visual features of icons, images and GUIs. Only two of the measures (the number of quadtree blocks and number of edge pixels, (Forsythe et al., 2003)) were tested on icons elsewhere; the rest has been introduced in this paper.

4 CONCLUSION

4.1 Conclusion

This thesis expands the knowledge on GUI visual aesthetics. It suggests computational method for predicting the aesthetics of GUIs, which is summarized in the predictive model of GUI aesthetics. The method has been tested in user studies and explained a substantial part of variance in the user scores of aesthetics. The method performed similarly (correlated with user scores) for different stimuli (webpages, mobile apps, app icons), which confirmed the validity and reliability. A number of implications follow from the work, which future research can build on.

4.1.1 Summary

The thesis started from the processing-fluency theory and incrementally developed it in the predictive model of GUI aesthetics. This work consisted of two phases: model development and model application. Chapters 2 described the model-development phase that explored visual complexity as an aesthetics predictor and outlined the complexity-based dimensions of design. The processing-fluency theory states that aesthetics stems from simplicity, at least in part (Reber et al., 2004). The first step included surveying psychological literature for the visual complexity determinants – often described for the sets of polygons instead of GUIs – and operationalized them for GUIs. Chapter 2 also classified such determinants of visual complexity in three groups describing the amount, organization and discriminability of visual detail. HCI researchers could frame their research using the classification or explore a determinant not operationalized in this thesis. A series of studies validated the complexity-aesthetics link on webpages and mobile apps, and thus, supported the main prediction of processing-fluency theory of aesthetics. The studies also assembled the ground-truth datasets of webpages and mobile apps. The datasets were used to match computed scores against user scores, and thus, to evaluate the performance of the computational methods. Further, the studies relied on crowdsourcing to minimize biases in stimuli sampling: such a crowdsourcing-based sampling method could replace the bias-prone practices of researchers sampling stimuli themselves.

The model-application phase of the work applied different aspects of the aesthetics model in different contexts. The study reported in chapter 3.1 explored the aspects of holistic aesthetics and was motivated by the recognition that the aesthetics of screenshots might differ from the holistic aesthetics of live GUIs. Several design aspects could differentiate such holistic aesthetics from screenshot-based aesthetics. For example, holistic aesthetics might be based on the aggregated impact of many pages of a website. The results showed that all webpages of the same website tended to have similar aesthetics and homepage-only impressions could approximate holistic website impression.

Visual diversity – which describes the visual differences amongst different layouts of a single GUI, such as the webpages of website – could be a prominent factor of holistic aesthetics. The high amount of diversity (e.g., between-page differences in a website) could either overload the user, and thus, decrease aesthetics, or engage the user with new content, and thus, increase aesthetics. Chapter 3.2 applied the computation-based methodology from Chapter 2 to explore GUI visual diversity. The exploration revealed a positive linear diversity-aesthetics relationship for both websites and mobile apps. Such finding supported the inclusion of diversity in aesthetics explorations (cf., Lavie & Tractinsky, 2004; Moshagen & Thielsch, 2010) and models.

Chapter 3.3 described the last step of the model-application phase, which consisted of a large-scale study of mobile app popularity on Google Play. The study demonstrated the ability of low-level visual features to predict the success of information systems, at least in the context of discretionary system use (cf., Diefenbach & Hassenzahl, 2011). The results revealed a significant correlation between the visual complexity and saliency of app icons and app popularity. Such correlation suggested the tendency of users to install the apps that they noticed and liked. Thus, the study demonstrated how the computational methods from the thesis could explain user behavior. This contribution concluded the work of this thesis.

4.1.2 Model of Design Aesthetics

The thesis has described the development and application of a predictive model of design aesthetics. The development phase of the work (Chapter 2) included building the model (Figure 39) and validating it in ten user studies on three types of designs. The validation resulted in a set of model parameters that customized the model for each design type. The prediction of aesthetics includes several actions: a design serves as the input in the model; eight design dimensions are estimated; the estimates are weighted and combined in an aesthetics score. The customization of model for various design types occurs in the parameters of algorithms that estimate design dimensions and in the vector of weights applied to the estimates.

The need for a customizable component became evident after comparing and contrasting the results of studies of chapter 2. Such analysis suggested that a single grand model for *all* design types would be ineffective at predicting aesthetics: all eight of design dimensions did influence the user impression for all design types, but such influences differed significantly amongst the types. Such differences appeared to stem from the design-specific aspects, such as the small available screen space of mobile apps or specific design guidelines for iPhones (cf., Chapter 2.3). The design-specific aspects might change the user perception of a design dimension, for example, a mobile-app user might not be concerned with the amount of white space if all apps occupy the whole screen and leave no white space, unlike webpages in a web browser. Design restrictions might also limit the variance in design-dimension estimates. Such limited variance limits the chances for covariance, and thus, decreases the strength of correlation between aesthetics and dimensions. For example, mobile apps might be optimized for finger touch, which could lead to spread-out content, and thus, to lower contour congestion for all apps.

The research of this thesis suggests that the customized model of aesthetics should be used for different design types: a design-specific set of algorithm parameters and dimension weights should be plugged in the generic model (Figure 39), making the model effective at predicting aesthetics for the specific design type.

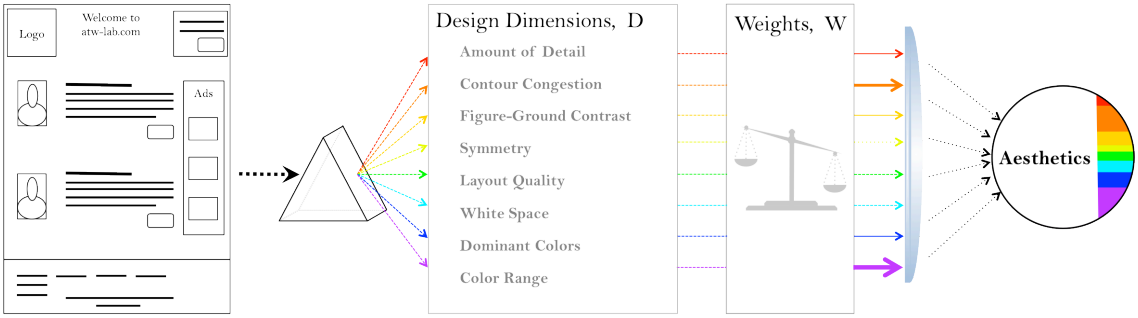


Figure 39. Modeling the aesthetics of GUI. A screenshot of GUI is split in design dimensions; the estimates of dimensions are weighted and recombined in an aesthetics score.

4.1.3 Future Work

Future research may build on the theoretical, practical and methodological contributions of this thesis. It could also address several limitations of the work and extend the predictive model of GUI aesthetics.

4.1.3.1 Theory

The work of this thesis may influence future research on aesthetics. The major influence would consist in outlining and testing a number of theoretical design dimensions. Future research could build on the outlined dimensions, and re-use one or several of them or expand the dimension list. For example, research could explore if the concepts of *metaphor* or *isolation* (Ramachandran & Seckel, 2012) would apply to GUI, and thus, should be added to the list as dimensions. Alternatively, future research could focus on a specific dimension to explore designs, instead of focusing on aesthetics as a whole. For example, researchers could explore font types and sizes as the major sources of contour congestion in designs. A proven font-congestion link would imply a font-aesthetics link.

The thesis also validated and expanded past theories and empirical findings. The examination of colorfulness suggested that future research should treat it as a two-factor design dimension consisting of color range and dominant colors. These two aspects differ in their impact on aesthetics: more dominant colors decreases aesthetics, whereas higher color range increases aesthetics. The difference might explain the empirical evidence of only a weak correlation between aesthetics and single-factor measure of colorfulness (Reinecke et al., 2013). The examination of the complexity-aesthetics link confirmed the main hypothesis of the processing fluency theory (Reber et al., 2004) and suggested the link could be exploited in future research. The examination of visual diversity confirmed that diversity was a component of GUI aesthetics (Moshagen & Thielsch, 2010) and suggested using it in the future studies on aesthetics. Finally, the analysis of stages of aesthetics impression (Thielsch et al., 2013) confirmed that the immediate (200ms exposure) and deliberate impression (4s exposure) carry over in the post-use impression. Such observation implies that future research could rely on the immediate or deliberate impressions instead of post-use impression, and still be able to generalize its results to use situations. Relying on the immediate and deliberate impressions could speed up future studies, as collecting the post-use impressions requires a significant effort – the user needs to interact with each system, not just see it.

4.1.3.2 Practice

The thesis has presented the computational method to estimate such aspects of design as visual complexity, saliency, diversity and aesthetics. A number of practical applications of the methods could be envisaged. Future research could use them instead of or in addition to the human-based measures, and thus, reduce the cost of conducting user studies.

The automatic evaluation makes possible processing designs on a large scale. The evaluation algorithms can sift through countless interfaces (e.g., mobile apps in Google Play) and detect the examples of very good design. Large-scale content providers (like, Google Play) could incorporate the methods in their content-ranking systems, and thus, increase the quality of content they deliver to the user.

A major application of the developed methods could be aiding design. However, such application will require the development of effective visualizations of design dimensions, since visualizations appear even more effective in aiding design than a label (e.g., "good design", "bad design"), Rosenholtz et al. (Rosenholtz et al., 2011). Chapter 3 featured several examples of such visualizations, but they may need to be developed further and tested in a study to evaluate their effectiveness in informing design. Teaching design could be another application, but may also require further research. Studies have shown that novice designer evaluators benefited from structured guidance (Lanzilotti et al., 2011). Future research could develop the structured-guidance material from the descriptions and visualizations of design dimensions that were used in the model. The material then could be used in class. The students would learn about the design dimensions or evaluate real designs in class.

4.1.3.3 Methodology

As a methodological contribution, this thesis has successfully demonstrated the application of the theory-led approach to model development, and thus, set an example for future work. The theory-led approach results in the models that have several advantages over the models from other approaches. First, the theory-led models rely on a real phenomenon (e.g., processing fluency) to define design dimensions rather than relying on researchers' intuition. Second, the models leverage only few dimensions that are unambiguously related to the outcome rather than leveraging multiple dimensions vaguely related to the outcome. Last, the models require no user input to estimate the aesthetics of new designs.

4.1.3.4 Limitations

Future research could extend the work of this thesis by addressing its several limitations. First, future work could operationalize and test several additional design dimensions. For example, this thesis has outlined in Chapter 2.1, but not operationalized the dimension of prototypicality. The resemblance of an interface to a prototype could increase interface aesthetics since user's expectations would be met (Roth et al., 2010). Future work could learn the appearance of the prototypical interface (cf., Kumar et al., 2013), and calculate the distance between an interface appearance and the prototype.

The comparison and contrasting of study results that are reported in the thesis has suggested that the predictive model of aesthetics should be customized for each design type. The types of design that the thesis explored included webpages, mobile apps and mobile app icons, which was not exhaustive. Many more design types exist, such as digital watch apps, videogames, or online advertisement. Future studies should detail and classify design types and genres, and develop the predictive models of aesthetics specifically for them, as this will substantially improve model performance.

Finally, the thesis investigated GUI visual diversity – the amount of visual change across different layouts of one GUI – but not the diversity-related concept of GUI dynamics. Similar to diversity, GUI dynamics corresponds to a visual change, but within a single GUI, such as an animated ad banner or unfolding menu on a webpage. GUI dynamics could be liked or disliked (Huhtala et al., 2011; Tractinsky et al., 2011) and could influence the user perception of GUI aesthetics, which future work should determine.

BIBLIOGRAPHY

Aksentijevic, A. & Gibson, K., 2012. Complexity equals change. Cognitive systems research, pp.1-16.

Altaboli, A. & Lin, Y., 2011. Investigating effects of screen layout elements on interface and screen design aesthetics. Advances in Human-Computer Interaction, 5, pp.1-10.

Archambault, D. & Purchase, H.C., 2013. Mental map preservation helps user orientation in dynamic graphs. Graph Drawing, pp.475-86.

Armstrong, T. & Detweiler-Bedel, B., 2008. Beauty as an emotion: the exhilarating prospect of mastering a challenging world. Review of general psychology, 12(4), pp.305-29.

Arnheim, R., 1954. *Art and Visual Perception*. Berkeley, CA: University of California Press.

Balinsky, H., 2006. Evaluating interface aesthetics: measure of symmetry. In *Digital Publishing Conference*. San Jose, 2006. International Society for Optics and Photonics.

Balinsky, H.Y..W.A.J..&.R.M.C., 2009. Aesthetic measure of alignment and regularity. In *the 9th ACM symposium on Document Engineering*. Munich, 2009. ACM.

Bauerly, M. & 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.*, 2006.

Beck, F., Burch, M. & Diehl, S., 2009. Towards an aesthetic dimensions framework for dynamic graph visualisations. In *13th International Conference Information Visualization.*, 2009. IEEE.

Beldad, A., De Jong, M. & Steehouder, M., 2010. How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in Human Behavior*, 26(5), pp.857-69.

Berlyne, D.E., 1971. *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Bhattacharya, S., Sukthankar, R. & Shah, M., 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *the International Conference on Multimedia.*, 2010. ACM.

Black, P.E., Scarfone, K. & Souppaya, M., 2008. Cyber security metrics and measures. In *Wiley Handbook of Science and Technology for Homeland Security*. John Wiley & Sons, Inc. pp.1-15.

Blythe, M. et al., 2010. ritical dialogue: interaction, experience and cultural theory. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems.*, 2010. ACM.

Böhmer, M. et al., 2011. Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *the 13th international conference on Human computer interaction with mobile devices and services.*, 2011. ACM.

Braddy, P.W., Meade, A.W. & Kroustalis, C.M., 2008. Online recruiting: The effects of organizational familiarity, website usability, and website attractiveness on viewers' impressions of organizations. *Computers in Human Behavior*, 24(6), pp.2992-3001.

Bravo, M.J. & Farid, H., 2004. Search for a category target in clutter.. *Journal of Perception*, 33(6), pp.643-52.

Cao, J., Mao, B. & Luo, J., 2010. A segmentation method for web page analysis using shrinking and dividing. *International Journal of Parallel, Emergent and Distributed Systems*, 25(2), pp.93-104.

Chae, M. & Kim, J., 2004. Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & Information Technology*, 23(3), pp.165-81.

Chang, D., Dooley, L. & Tuovinen, J.E., 2002. Gestalt Theory in Visual Screen Design — A New Look at an old subject. In *the 7th World Conference on Computers in Education*. Melbourne, 2002.

Chatzichristofis, S.A. & Boutalis, Y.S., 2008. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *Ninth International Workshop on Image Analysis for Multimedia Interactive Services.*, 2008. IEEE.

Chen, G. & Choi, B., 2008. Web page genre classification. In *the 2008 ACM symposium on Applied computing.*, 2008. ACM.

Chittaro, L., 2011. Designing visual user interfaces for mobile applications. In *3rd ACM SIGCHI symposium on Engineering interactive computing systems.*, 2011. ACM.

Choi, J.H. & Lee, H.J., 2012. Facets of simplicity for the smartphone interface: A structural model. *International Journal of Human-Computer Studies*, 70(2), pp.129-42.

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

Datta, R., Joshi, D., Li, J. & Wang, J.Z., 2006. Studying aesthetics in photographic images using a computational approach. In *9th European Conference on Computer Vision.*, 2006. Springer Berlin Heidelberg.

De Angeli, A., Hartmann, J. & Sutcliffe, A., 2008. The effect of brand on the evaluation of websites. In *INTERACT.*, 2008. Springer Berlin Heidelberg.

De Angeli, A., Sutcliffe, A. & Hartmann, J., 2006. Interaction, usability and aesthetics: what influences users' preferences? In *the 6th conference on Designing Interactive systems*., 2006. ACM.

Deng, L. & Poole, M.S., 2010. Affect in Web Interfaces: A Study of the Impacts of Web Page Visual Complexity and Order. *Mis Quarterly*, 34(4), pp.711-30.

Diefenbach, S. & Hassenzahl, M., 2011. The dilemma of the hedonic–Appreciated, but hard to justify. *Interacting with Computers*, 23(5), pp.461-72.

Djamasbi, S., Sigel, M. & Tullis, T., 2011. Visual heirarchy and viewing behavior: an eye tracking study. *Human-computer interaction, design and development approaches, Lecture notes in computer science*, pp.331-40.

Donderi, D.C., 2006. Visual complexity: a review. *Psychological bulletin*, 132(1), pp.73-97.

Duncan, J. & Humphreys, G.W., 1989. Visual search and stimulus similarity. *Psychological review*, 96(3), pp.433-58.

Duncan, J., 1984. Selective attention and the organization of visual information. *Journal of experimental psychology*, 113(4), pp.501-17.

Fei-Fei, L., Iyer, A., Koch, C.. & Perona, P., 2007. What do we perceive in a glance of a real-world scene? *Journal of vision*, 7(1), pp.1-29.

Forsythe, A., 2009. Visual Complexity: Is That All There Is? In D. Harris, ed. *Engineering Psychology and Cognitive Ergonomics*. Springer Berlin Heidelberg. pp.158-66.

Forsythe, A., Sheehy, N. & Sawey, M., 2003. Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computers*, 35(2), pp.334-42.

Gartus, A. & Leder, H., 2013. The small step toward asymmetry: Aesthetic judgment of broken symmetries. *i-Perception*, pp.352-55.

Grudin, J., 1989. The case against user interface consistency. *Communications of the ACM*, 32(10), pp.1164-73.

Gwizdka, J. & Spence, I., 2006. What can searching behavior tell us about the difficulty of information tasks? A study of Web navigation. *Proceedings of the American Society for Information Science and Technology*, 43(1), pp.1-22.

Hall, R.H. & Hanna, P., 2004. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23(3), pp.183-95.

Harper, S., Jay, C., Michailidou, E. & Quan, H., 2013. Analysing the visual complexity of web pages using document structure. *Behavious & Information Technology*, 32(5), pp.491-502.

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

Harrington, S.J. et al., 2004. Aesthetic measures for automated document layout. *the 2004 ACM symposium on Document engineering*, October. pp.109-11.

Hartmann, J., De Angeli, A. & Sutcliffe, A., 2008. Framing the user experience: information biases on website quality judgement. In *CHI*., 2008. ACM.

Hartmann, J., Sutcliffe, A. & De Angeli, A., 2007. Inverstigating attractiveness in web user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*., 2007. ACM.

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

Hassenzahl, M. & Monk, A., 2010. The inference of perceived usability from beauty. *Human-Computer Interaction*, 25(3), pp.235-60.

Hassenzahl, M., 2004. The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction*, 19(4), pp.319-49.

Hassenzahl, M., 2005. The thing and I: understanding the relationship between user and product. *Funology*, pp.31-42.

Hassenzahl, M., 2012. Commentary on: Tractinsky, Noam (2012): Visual Aesthetics: in human-computer interaction and interaction design. In M. Soegaard & R.F. Dam, eds. *Encyclopedia of Human-Computer Interaction*. The Interaction-Design.org Foundation.

Hekkert, P., Snelders, D. & 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), pp.111-24.

Huhtala, J. et al., 2011. Animated UI transitions and perception of time: a user study on animated effects on a mobile screen. In *SIGCHI Conference on Human Factors in Computing Systems*., 2011. ACM.

Iakovidou, C. et al., 2014. Searching images with MPEG-7 (& mpeg-7-like) powered localized descriptors: the SIMPLE answer to effective content based image retrieval. In *Content-Based Multimedia Indexing (CBMI)*., 2014. ACM.

Itti, L., Koch, C. & Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 20(11), pp.1254-59.

Ivory, M.Y. & Hearst, M.A., 2002. Statistical profiles of highly-rated web sites. In *SIGCHI conference on Human factors in computing systems*., 2002. ACM.

Ivory, M.Y., Sinha, R.R. & 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.

Kaplan, S., 1973. Cognitive maps, human needs and the designed environment. *Environmental Design Research: Selected papers*, pp.275-83.

Karapanos, E., Zimmerman, J., Forlizzi, J. & Martens, J.-F., 2009. User experience over time: an initial framework. In *the 27th international conference on Human factors in computing systems*. Boston, 2009. ACM.

Khalid, H., Shihab, E., Nagappan, M. & Hassan, A., 2014. What do mobile app users complain about? *Software, IEEE*, 32(3), pp.70-77.

Kim, H. & Fesenmaier, D.R., 2008. Persuasive design of destination web sites: An analysis of first impression. *Journal of Travel Research.*, 47(1), pp.3-13.

Kim, H.W., Lee, H.L. & Son, J.E., 2011. An exploratory study on the determinants of smartphone app purchase. In *The 11th International DSI and the 16th APDSI Joint Meeting*. Taipei, 2011.

Kim, J., Lee, J. & Choi, D., 2003. Designing emotionally evocative homepages: an empirical study of the quantitative relations between design factors and emotional dimensions. *Internation Journal of Human Computer Studies*, 59(6), pp.899-940.

Kootstra, G., de Boer, B. & Schomaker, L.R.B., 2011. Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive computation*, pp.223-40.

Kumar, R. et al., 2013. Webzeitgeist: design mining the web. In *the SIGCHI Conference on Human Factors in Computing Systems.*, 2013. ACM.

Kurosu, M. & Kashimura, K., 1995. pparent usability vs. inherent usability: experimental analysis on the determinants of the apparent usability. In *Human factors in computing systems.*, 1995. ACM.

Lanzilotti, R., Ardito, C., Costabile, M.F. & De Angeli, A., 2011. Do patterns help novice evaluators? A comparative study. *International journal of human-computer studies*, 69(1), pp.52-69.

Lavie, T.. & Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies*, 60(3), pp.269-98.

Lee, S. & Koubek, R.J., 2010. Understanding user preferences based on usability and aesthetics before and after actual use. *Interacting with Computers*, 22(6), pp.530-43.

Leuthold, S. et al., n.d. Vertical versus dynamic menus on the world wide web: Eye tracking study measuring the influence of menu design and task complexity on user performance and subjective preference. *Computers in human behavior*, 27(1), pp.459-72.

Levi, D.M., 2008. Crowding—An essential bottleneck for object recognition: A mini-review. *Vision Research*, 48(5), pp.635-54.

Lindgaard, G. et al., 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *Transactions on computer-human interaction*, 18(1), pp.1-30.

Lindgaard, G., Fernandes, G., Dudek, C. & Brown, J., 2006. Attention web designers: You have 50 milliseconds to make a good first impressoin! *Behavior & information technology*, 25(2), pp.115-26.

Ling, C., Hwang, W. & Salvendy, G., 2007. A survey of what customers want in a cell phone design. *Behaviour & Information Technology*, 26(2), pp.149-63.

Loy, G. & Eklundh, J.O., 2006. Detecting symmetry and symmetric constellations of features. In Leonardis, A., Bischof, H. & Pinz, A. *Computer Vision – ECCV 2006*. Springer Berlin Heidelberg. pp.508-21.

Lux, M. & Chatzichristofis, S.A., 2008. Lire: lucene image retrieval: an extensible java cbir library. In *16th ACM international conference on Multimedia.*, 2008. ACM.

Machilsen, B., Pauwels, M. & Wagemans, J., 2009. The role of vertical mirror symmetry in visual shape detection. *Journal of Vision*, 9(12), pp.1-11.

Machilsen, B., Pauwels, M. & Wagemans, J., 2009. The role of vertical mirror symmetry in visual shape detection. *Journal of Vision*, 9(12), pp.1-11.

Martindale, C., Moore, K. & Borkum, J., 1990. Aesthetic preference: Anomalous findings for Berlyne's psychobiological theory. *The American Journal of Psychology*, pp.53-80.

McCay-Peet, L., Lalmas, M. & Navalpakkam, V., 2012. On saliency, affect and focused attention. In *SIGCHI Conference on Human Factors in Computing Systems.*, 2012. ACM.

McDougall, S. & Reppa, I., 2013. Ease of icon processing can predict icon appeal. In *15th International Conference, HCI International.*, 2013. Springer Berlin Heidelberg.

McDougall, S.J., de Bruijn, O. & Curry, M.B., 2000. Exploring the effects of icon characteristics on user performance: the role of icon concreteness, complexity, and distinctiveness. *Journal of Experimental Psychology: Applied*, 6(4), pp.291-306.

Michailidou, E., Harper, S. & Bechhofer, S., 2008. Visual complexity and aesthetic perception of web pages. In *the 26th annual ACM international conference on Design of communication*. Lisbon, 2008. ACM.

Miniukovich, A. & De Angeli, A., 2014a. Quantification of Interface Visual Complexity. In *the 2014 International Working Conference on Advanced Visual Interfaces*. Como, 2014a. ACM.

Miniukovich, A. & De Angeli, A., 2014b. Visual Impression of Mobile App Interfaces. In *NordiCHI'14.*, 2014b. ACM.

Miniukovich, A. & De Angeli, A., 2015a. Computation of Interface Aesthetics. In *CHI'15.* Seoul, 2015a.

Miniukovich, A. & De Angeli, A., 2015b. Visual diversity and user interface quality. In *the 2015 British HCI Conference.*, 2015b. ACM.

Miniukovich, A. & De Angeli, A., 2016. Pick Me! Getting Noticed on Google Play. In *CHI'16.*, 2016. ACM.

Moore, R.S., Stammerjohan, C.A. & Coulter, R.A., 2005. Banner advertiser-web site context congruity and color effects on attention and attitudes. *Journal of Advertising*, 34(2), pp.71–84.

Mormann, M.M., Navalpakkam, V., Koch, C.. & Rangel, A., 2012. Relative visual saliency differences induce sizable bias in consumer choice. *Journal of Consumer Psychology*, 22(1), pp.67–74.

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

Moshagen, M., Musch, J. & Göritz, A.S., 2009. A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, 52(10), pp.1311–20.

Mudambi, S.M. & Schuff, D., 2010. What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quarterly*, pp.185–200.

Nadkarni, S. & Gupta, R., 2007. A task-based model of perceived website complexity. *Mis Quarterly*, pp.501–24.

Nadkarni, S. & Gupta, R., 2007. A Task-Based Model of Perceived Website Complexity. *Mis Quarterly*, 31(3).

Navalpakkam, V. & Itti, L., 2005. Modeling the influence of task on attention. *Vision Research*, 45(2), pp.205–31.

Ng, A.W. & Chan, A.H., 2008. Visual and cognitive features on icon effectiveness. In *international multiconference of engineers and computer scientists.*, 2008.

Ngo, D.C.L., 2001. Measuring the aesthetic elements of screen designs. Displays. *Displays*, 22(3), pp.73–78.

Ngo, D.C.L., Teo, L.S. & Byrne, J.G., 2000. Formalising guidelines for the design of screen layouts. *Dislpays*, 21(1), pp.3–15.

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

Oliva, A., Mack, M.L., Shrestha, M. & Peeper, A., 2004. Identifying the perceptual dimensions of visual complexity of scenes. *the 26th Annual Meeting of the Cognitive Science Society*, August.

Park, S.E., Choi, D.. & Kim, J., 2004. Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users. *Interacting with Computers*, 16(2), pp.351–76.

Parkhurst, D.J. & Niebur, E., 2004. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19(3), pp.783–89.

Porat, T. & Tractinsky, N., 2012. It's a Pleasure Buying Here: The Effects of Web-Store Design on Consumers' Emotions and Attitudes. *Human–Computer Interaction*, 27(3), pp.235–76.

Pothos, E.M. & Ward, R., 2000. Symmetry, repetition, and figural goodness: An investigation of the weight of evidence theory. *Cognition*, 75(3), pp.B65–78.

Purchase, H.C., Freeman, E. & Hamer, J., 2012. An exploration of visual complexity. *Digrammatic representation and inferences*, pp.200–13.

Purchase, H.C., Hamer, J., Jameson, A. & Ryan, O., 2011. Incesitigating objective measures of web page aesthetics and usability. In *12th Australasian user interface conference (AUIC 2011).* Perth, 2011. Australian computer society, Inc.

Rafaeli, A. & Vilnai-Yavetz, I., 2004. Instrumentality, aestheitcs and symdolism of hpysical artifacts as triggers of emotion. *Theoretical Issues in Ergonomics Science*, 5(1), pp.91–112.

Ramachandran, V.S. & Hirstein, W., 1999. The science of art: A neurological theory of aesthetic experience. *The journal of consciousness studies*, 6, pp.15–35.

Ramachandran, V.S. & Seckel, E., 2012. Neurology of Visual Aesthetics: Indian Nymphs, Modern Art, and Sexy Beaks. In P. Shimamura & S.E. Palmer, eds. *Aesthetic Science.* Oxford University Presss. pp.375–90.

Reber, R., 2012. Processing fluency, aesthetic pleasure, and culturally shared taste. In Shimamura, A.P. & Palmer, S.E. *Aesthetic Science: Connecting Minds, Brains, and Experience.* Oxford: Oxford University Press. pp.223–49.

Reber, R., Schwarz, N. & Winkielman, P., 2004. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4), pp.364–82.

Reber, R., Winkielman, P. & Schwarz, N., 1999. Effects of perceptual fluency on affective judgments. *Psychological Science*, 9(1), pp.45–48.

Reber, R., Wurtz, P. & Zimmermann, T.D., 2003. Exploring "fringe" consciousness: The subjective experience of perceptual fluency and its objective bases. *Consciousness and Cognition*, 13(1), pp.47–60.

Reinecke, K. & Gajos, K.Z., 2014. Quantifying visual preferences around the world. In *the 32nd annual ACM conference on Human factors in computing systems.*, 2014. ACM.

Reinecke, K. et al., 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *CHI*. Paris, 2013. ACM.

Reppa, I. & McDougall, S., 2015. When the going gets tough the beautiful get going: aesthetic appeal facilitates task performance. *Psychonomic bulletin & review*, 22, pp.1243-54.

Riva, A.D. et al., 2010. Two new aesthetic measures for item alignment. In *the 10th ACM symposium on Document engineering*. Manchester, 2010. ACM.

Rizzi, A. & McCann, J.J., 2007. On the behavior of spatial models of color. In *SPIE 6493, Color Imaging XII: Processing, Hardcopy, and Applications.*, 2007. International Society for Optics and Photonics.

Robertson, G., Czerwinski, M., Fisher, D. & Lee, B., 2009. Selected human factors issues in information visualization. *Reviews of human factors and ergonomics*, 5(1), pp.41-81.

Rosenholtz, R., Dorai, A. & Freeman, R., 2011. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)*, 8(2).

Rosenholtz, R., Li, Y. & Nakano, L., 2007. Measuring visual clutter. *Journal of vision*, 7(2), pp.1-22.

Roth, S.P. et al., 2010. Mental models for web objects: Where do users expect to find the most frequent objects in online shops, news portals, and company web pages? *Interacting with Computers*, 22(2), pp.140-52.

Santayana, G., 1955. *The sense of beauty: Being the outline of aesthetic theory*. Courier Corporation.

Sereno, S.C. & Rayner, K., 2003. Measuring word recognition in reading: eye movements and event-related potentials. *Trends in cognitive sciences*, 7(11), pp.489-93.

Shimamura, A.P., 2012. Towards a Science of Aesthetics: Issues and Ideas. In A.P. Shimamura & P.S. E., eds. *Aesthetic Science: Connecting Minds, Brains, and Experience*. Oxford. pp.3-30.

Silvia, P.J., 2012. Human emotions and aesthetic experience: An overview of empirical aesthetics. In A.P. Shimamura & S.E. Palmer, eds. *Aesthetic science: Connecting minds, brains, and experience*. Oxford University Press. pp.250-75.

Smith-Gratto, K. & Fisher, M.M., 1998-99. Gestalt theory: a foundation for instructional screen design. *Journal of Educational Technology Systems*, 27(4), pp.361-72.

Sonderegger, A. & Sauer, J., 2010. The influence of design aesthetics in usability testing: Effects on user performance and perceived usability. *Applied Ergonomics*, 41(3), pp.403-10.

Sonderegger, A., Zbinden, G., Uebelbacher, A. & Sauer, J., 2012. The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. *Ergonomics*, 55, pp.713-30.

Sutcliffe, A., 2002. Assessing the reliability of heuristic evaluation for Web site attractiveness and usability. In *the 35th Hawaii International Conference on System Sciences.*, 2002. IEEE.

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

Tabachnick, B. & Fidell, L., 2007. *Using Multivariate Statistics*. 5th ed. Pearson Education Inc.

Tamura, H., Mori, S. & Yamawaki, T., 1978. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6), pp.460-73.

Thielsch, M.T. & Hirschfeld, G., 2010. High and low spatial frequencies in website evaluations. *Ergonomics*, 53(8), pp.972-78.

Thielsch, M.T. & Hirschfeld, G., 2012. Spatial frequencies in aesthetic website evaluations–explaining how ultra-rapid evaluations are formed. *Ergonomics*, 55(7), pp.731-42.

Thielsch, M.T., Blotenberg, I. & Jaron, R., 2013. User evaluation of websites: From first impression to recommendation. *Interacting with Computers*, 26(1), pp.89-102.

Tinio, P.P. & Leder, H., 2009. Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica*, 130(3), pp.241-50.

Tractinsky, N. & Zmiri, D., 2006. Exploring attributes of skins as potential antecedents of emotion in HCI. *Aesthetic Computing*, pp.405-22.

Tractinsky, N., 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In *the ACM SIGCHI Conference on Human factors in computing systems.*, 1997. ACM.

Tractinsky, N., 2013. Visual Aesthetics. In M. Soegaard & R.F. Dam, eds. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Aarhus: The Interaction-Design.org Foundation.

Tractinsky, N., Cokhavi, A., Kirschenbaum, M. & Sharfi, T., 2006. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International journal of human-computer studies*, 64(11), pp.1071-83.

Tractinsky, N., Inbar, O., Tsimhoni, O. & Seder, T., 2011. Slow down, you move too fast: Examining animation aesthetics to promote eco-driving. In *3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications.*, 2011. ACM.

Tractinsky, N., Katz, A.S. & Ikar, D., 2000. What is beautiful is usable. *Interacting with computers*, 13(2), pp.127-45.

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), pp.194–214.

Tuch, A.N. et al., 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, pp.794–811.

Tuch, A.N., Bargas-Avila, J.A. & Opwis, K., 2010. Symmetry and aesthetics in website design: It'sa man's business. *Computers in Human Behavior*, 26(6), pp.1831–37.

Tuch, A.N., Bargas-Avila, J.A., Opwis, K. & Wilhelm, F.H., 2009. Visual complexity of websites: effects on users' experience, physiology, perfomance, and memory. *International journal of human-computer studies*, 67, pp.703–15.

van den Berg, R., Cornelissen, F.W. & Roerdink, J.B., 2009. A crowding model of visual clutter. *Journal of Vision*, 9(4), pp.1–11.

van der Geest, T. & Loorbach, N., 2005. Testing the visual consistency of web sites. *Technical Communication*, 52(1), pp.27–36.

van der Helm, P.A., 2000. Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, pp.770–800.

van Leuken, R.H., Garcia, L., Olivares, X. & van Zwol, R., 2009. Visual diversification of image search results. In *18th international conference on World wide web.*, 2009. ACM.

Van Schaik, P. & Ling, J., 2008. Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. *Interacting with Computers*, 20(3), pp.419–32.

van Schaik, P. & Ling, J., 2009. The role of context in perceptions of the aesthetics of web pages over time. *International Journal of Human-Computer Studies*, 67(1), pp.79–89.

Wertheimer, M., 1938. Laws of Organization in Perceptual Forms (partial translation). In Ellis, W.B. *A Sourcebook of Gestalt Psycychology*. Harcourt Brace. pp.71–88.

Whitfield, T.W.A., 1983. Predicting preference for familiar, everyday objects: An experimental confrontation between two theories of aesthetic behaviour. *Journal of environmental psychology*, 3(3), pp.221–37.

Whittlesea, B.W. & 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), pp.141–65.

Winkielman, P., Schwarz, N., Fazendeiro, T.A. & Reber, R., 2002. The hedonic marking of processing fluency: Implications for evaluative judgment. In Musch, J. & Klauer, K.C. *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*. Psychology Press. pp.195–223.

Wolfe, J.M. & Horowitz, T.S., 2004. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6), pp.495–501.

Wong, N., Carpendale, S. & Greenberg, S., 2003. EdgeLens: An Interactive Method for Managing Edge Congestion in Graphs. *IEEE Symposium on Information Visualization*, 19-21 October. pp.51–58.

Wright, P., Wallace, J. & McCarthy, J., 2008. Aesthetics and experience-centered design. *CM Transactions on Computer-Human Interaction (TOCHI)*, 15(4).

Wu, O., Chen, Y., Li, B. & 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.*, 2011. ACM.

Wu, O., Hu, W. & Shi, L., 2013. Measuring the visual complexities of Web pages. *ACM Transactions on the Web (TWEB)*, 7(1), pp.1–34.

Yu, H. & Winkler, S., 2013. Image complexity and spatial information. In *Quality of Multimedia Experience (QoMEX).*, 2013. IEEE.

Zheng, X.S., Chakraborty, I., Lin, J.J.W. & 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.*, 2009. ACM.