Pick me! Getting Noticed on Google Play

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

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

ABSTRACT
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.

Author Keywords
Visual complexity; visual saliency; mobile apps; app icon quality; consumer choice; computational methods.

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

INTRODUCTION
Millions of apps populate Google Play[1]. 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 [44]. Many apps fail to impress.

This paper offers a new line of research. Instead of taking user complaints [11], menu structure [4] or app layout screenshots [24] 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 1) and mobile versions (Figure 2) of Google Play. Icons have become "*the visual expression of a brand's products, services, and tools*"[2] 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.

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

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

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858552

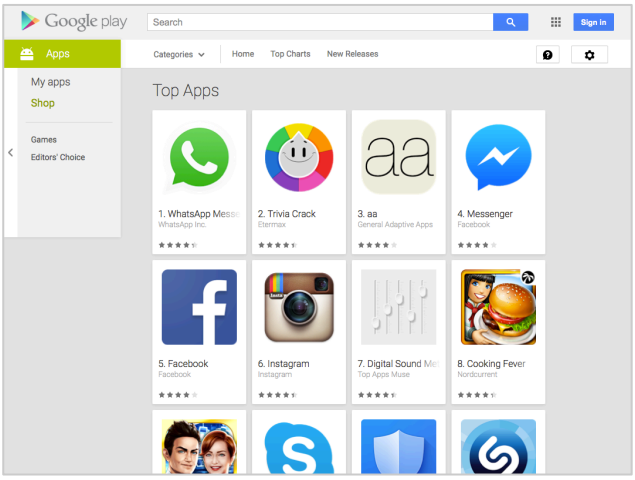

Figure 1. A listing of apps on Google Play as seen from a wide-screen device.

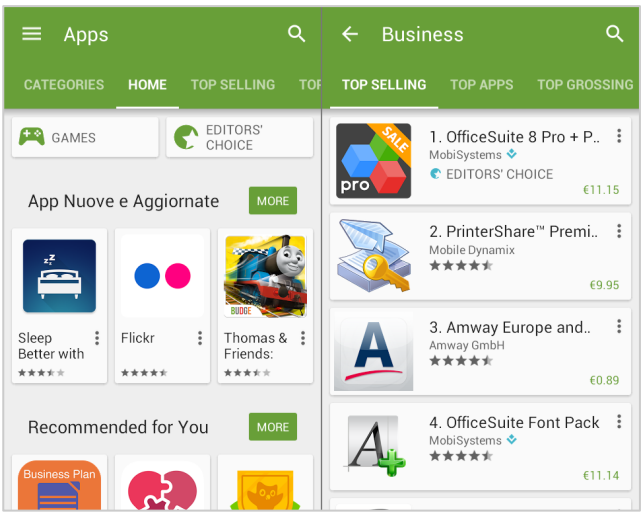

Figure 2. App listings on Google Play as seen from a mobile device.

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 [27] [13], and reviewed the types of user complaints showing their effect on app success [11]. Other researchers looked at individual apps, outside of marketplaces. They delved in app user behavior and found stable behavioral patterns [3], linked the structure of navigation menus and user satisfaction [4], and tried to computationally model app GUI appreciation using app screenshots [24]. 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.

First Impression

The models of user experience evolution distinguished three stages of UX [38]: 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 [16]. However, the appreciation particularly prevails during the first stage, orientation, when the user chooses to install and stick with an app [38].

The appreciation of visual aesthetics happens very quickly, in under a half second [15,23], changes little after it is formed [45], and affects user actions [40]. In a study on websites, Thielsch et al. [40] linked first impression to the overall impression, intention to revisit and intention to recommend to friends. Kim and Fesenmaier [12] 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 [19] and low complexity [34].

Visual Saliency and Complexity

Among the qualities that constitute a good icon (e.g., McDougal et al. [20] 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 [20]. Other visual features that also could contrast an icon from its surroundings include color, edge orientation, texture, size, motion and flicker [49]. 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 [19].

The other visual feature of icons, their visual complexity, was investigated more often than any other feature [30] and was even attempted to quantify automatically [7] [6]. Icon complexity corresponds to the amount and intricacy of detail within an icon (cf., [30]), and has been linked to a range of outcomes. For example, McDougal & Reppa [21]

linked icon visual complexity to visual appeal and search time. They explained the link via the concept of processing fluency [34]: 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. [47] 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. [31] 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 [23] 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.

Metrics and Measures

Both *metric* and *measure* refer to estimates and estimating. However, we follow the tradition [2] 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 [10] 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., [32,10]). 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., [9] 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 [28] 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. [50] 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 [22] 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. [33] 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 [47,50,6,9]). 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 [35,51,7] 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. [7] 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 [23] 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.

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.

Hypotheses

Past work suggested that mental attention-guidance mechanisms might largely reduce the input that humans process consciously [49,26]. The input might then be further reduced down to what is considered simple, and therefore, likable [15,34]. We applied these ideas in a schema of app selection, Figure 3. 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 3) 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*".

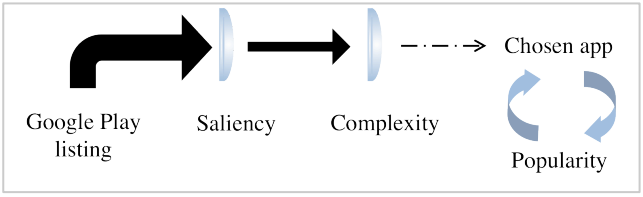

Figure 3. The schema of app selection: in the absence of top-down selection factors (e.g., task at hand, [49,29]), the bottom-up, vision-based factors determine the choice.

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 1, Figure 2). 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.

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).

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.

Saliency

We relied[3] on Itti et al.'s model of saliency-based visual attention [10]. 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

[3] We re-used Jonathan Harel's implementation of the saliency algorithm [10], which he kindly published online, http://www.vision.caltech.edu/~harel/share/gbvs.php

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 1, Figure 2), 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 4). 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 4, in the red square) was averaged across ten target-surround images and taken as the metric of saliency.

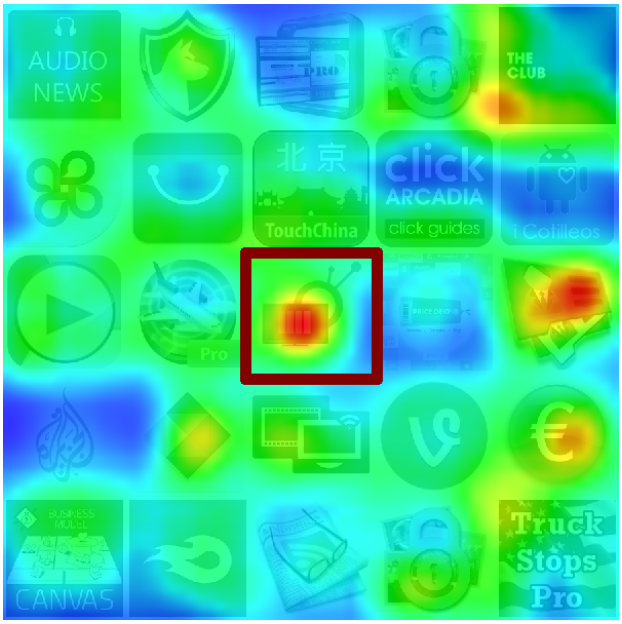

Figure 4. The target icon (in a red square) surrounded by 24 other icons served as an input for the saliency metric. The corresponding saliency map is overlaid as a heatmap.

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 (quadtree decomposition and high-pass filtering). We included the extra measures to link up our work with past work [6,7,41,42].

Color Variability

Color variability, consists of two aspects: number of dominant colors and color range [23]. 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 [23] and we re-applied the idea to icons We counted the number of dominant colors using the method of uniform color quantization (cf. [23,25,22]): 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 5 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 5. Color quantization. Color semi-tones and shades from two icons were discarded and only the main colors were counted.

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, [37]), Subband Entropy (an estimate of feature redundancy in an image, [37]), image file sizes in JPEG [47], and number of contour pixels [37]. The measures were observed to strongly intercorrelate [23,25]. We chose to compute the number of contour pixels: it was the simplest measure, which was also tested on icons [7]. Contours were detected using the Canny edge detector (low threshold - .11, high threshold - .27; cf., [37]). 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 6 shows the contours of an icon.

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 [8] demonstrated higher text-background contrast to decrease

the effort of reading. We used the measure of contrast from [23,25]. 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, [23]) 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 7 demonstrates the contrast measure: stronger edges are green, subtler edges are red.

Figure 6. Icon contour detection. Contour pixels are counted and taken as an estimate of amount of detail.

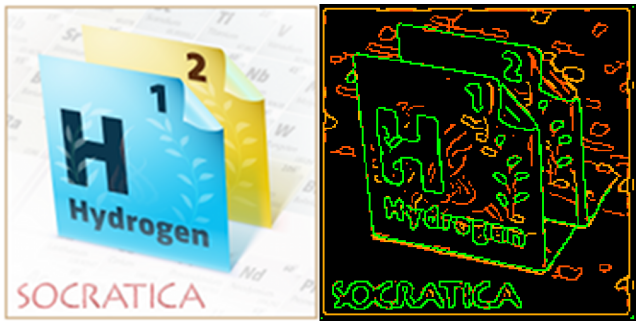

Figure 7. Contrast measurement. The transition from green to red reflects lowering contrast.

Symmetry

Vertical mirror symmetry facilitates shape perception [18] and has been shown to increase the appeal of patterns (black-white dots, lines and shapes, [43]) and webpages. For GUIs, Researchers developed several automatic block-based measures of symmetry [1,22]. 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[4] of detecting local symmetries from the object detection and recognition domain [17]. 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., [37]) – correlated with icon symmetry estimates by the method [17], $r(497) = .52$, $p < .001$. 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 [23]. 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. ([23] used 4-pixel radius areas, which we reduced to reflect a small icon size.) The ratio of symmetrical keypoints to all keypoints was taken as an estimate of icon symmetry. Figure 8 demonstrates the symmetry measure. Symmetrical keypoints are green; asymmetrical keypoints are red.

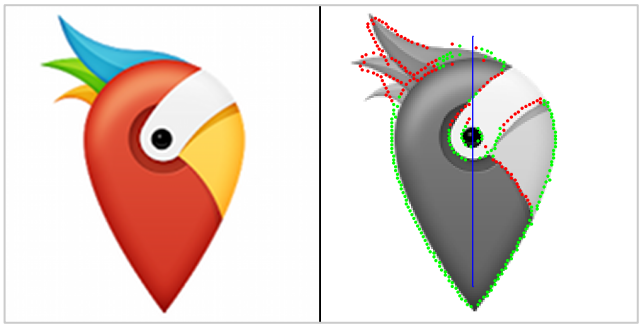

Figure 8. Symmetry measure. Symmetrical contour points (green) have a pair across the central vertical axis (the blue line). Non-symmetrical points (red) have no pair.

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 9). 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 9); the number of quadtree blocks strongly correlates with other measures of

[4] We used Loy et al.'s [17] implementation of the method, from http://www.nada.kth.se/~gareth/homepage/local_site/code.htm

image detail [7], 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 [7,51,35,50].

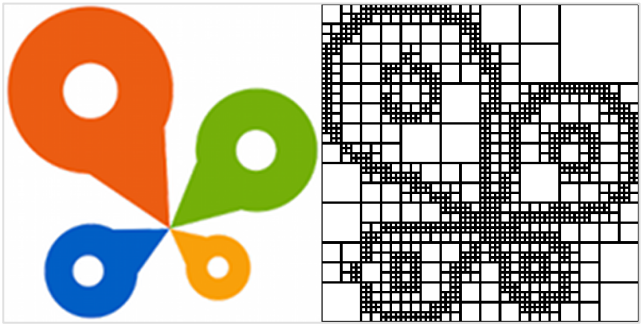

Figure 9. Quadtree decomposition. An image is split in square blocks till the pixel luminance within a block varies less than 25% of maximal luminance.

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 [6] and GUI aesthetics [42,41]. 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 10, center) from the original icons (Figure 10, left) to get the high-frequency information (Figure 10, right). We then considered two high-pass based measures: the number of non-zero pixels (all dark pixels, Figure 10 right) and the average luminance of non-zero pixels (cf., Yu et al.'s [13] 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 11); we titled it as contour energy.

Figure 10. Frequency filtering. The central image is processed with a low-pass filter; the right image - with a high-pass filter.

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 [48]. We used the contour congestion measure from [23]. 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 12 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., [37]). We normalized the congestion scores by the number of quadtree blocks, and thus, decoupled them from the set size.

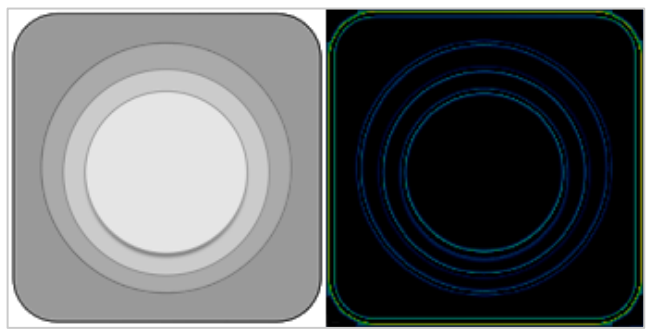

Figure 11. Contour energy. The brighter and greener contours of the processed icon (on the right) correspond to the larger visual contour-background difference of the original icon (on the left).

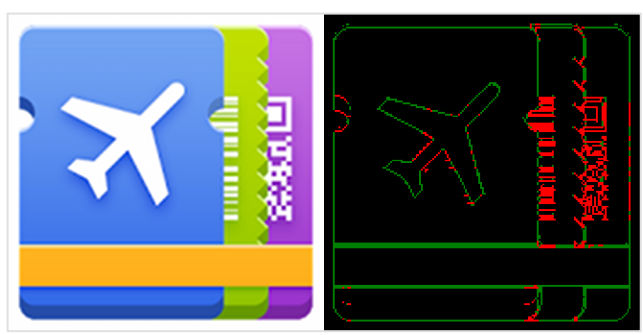

Figure 12. Contour congestion. Congested contours are marked red.

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.

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 13). 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.

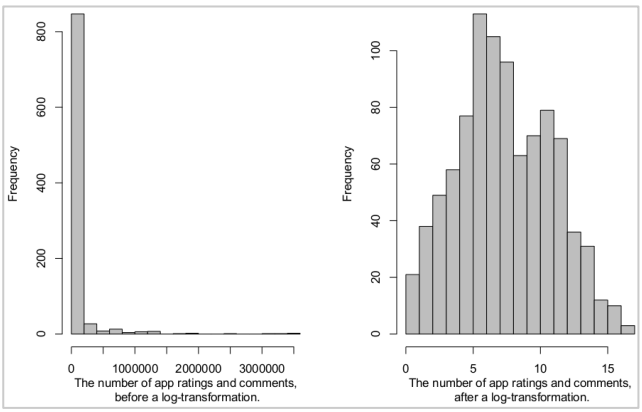

Figure 13. A review of the histogram of rating counts (left) shows the vast majority of apps had relatively few ratings; the data needed to be log-transformed (right).

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 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 1); 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 ($F(5,924) = 76.96$, $p < .001$, $\eta^2 = .29$) than on the mean app rating ($F(5,924) = 8.37$, $p < .001$, $\eta^2 = .04$). This supported hypothesis 3. Seven computed independent variables were put in a stepwise regression with backward exclusion to select the best linear model (Table 2). The inclusion of app genre – a categorical, non-computed variable – in the linear model further improved model fit ($R^{2)}$ up to .38, Table 2. 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 1. 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 2. A linear regression model of app popularity. (The outcome variable is the number of app ratings.)

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.

App Popularity

We hypothesized that users may chose apps because of their visually salient but simple icons design (cf., Figure 3). To test such a hypothesis, we calculated eight visual-complexity measures and a single saliency metric, and

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

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 2. 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 [11].

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 $\eta^2 = .29$), and little of app appreciation (the size of effect on the mean ratings was $\eta^2 = .04$).

Computational Method

To estimate image visual saliency, we extended the well-known method from Itti et al. [10]. 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 [10] 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 [10] 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 2). We might speculate visual saliency indeed selected candidate-objects for further, complexity-based mental processing (Figure 3, cf., [49]).

We estimated icon visual complexity with eight measures; all measure scores correlated with the number of app ratings (Table 1). 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., [37]). As expected (cf., [7]), the measures strongly cross-correlated and were included in the analysis as a link to the past work (namely, [7] [6]). 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 1). 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 ([46,23], 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.

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 [14] 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., [35]) or textures [39] 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 [5]. Finally, a future study may computationally address visual appeal – a quality shown to make an icon to stand out, particularly in stressful situations [36].

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 [11] or in-app visual consistency [24]. 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, [7]) were tested on icons elsewhere; the rest has been introduced in this paper.

ACKNOWLEDGEMENT

This paper research was partially funded by the Ministero dell'Istruzione, Universita e Ricerca, via the project "Città Educante", project code CTN01_00034_393801.

REFERENCES

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

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

3. Böhmer, M., Hecht, B., Schöning, J., Krüger, A., and Bauer, G. 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, 47-56.

4. Chae, M. and Kim, J. 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 (2004), 165-181.

5. De Angeli, A., Hartmann, J., and Sutcliffe, A. The effect of brand on the evaluation of websites. In *INTERACT* (2008), Springer Berlin Heidelberg, 638-651.

6. Forsythe, A. Visual Complexity: Is That All There Is? In Harris, Don, ed., *Engineering Psychology and Cognitive Ergonomics*. Springer Berlin Heidelberg, 2009.

7. Forsythe, A., Sheehy, N., and Sawey, M. Measuring icon complexity: An automated analysis. *Behavior Research Methods, Instruments, & Computers*, 35, 2 (2003), 334-342.

8. Hall, R. H. and Hanna, P. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. *Behaviour & information technology*, 23, 3 (2004), 183-195.

9. Harper, Simon, Jay, Caroline, Michailidou, Eleni, and Quan, Huangmao. Analysing the visual complexity of web pages using document structure. *Behavious & Information Technology*, 32, 5 (2013), 491-502.

10. Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 20, 11 (1998), 1254-1259.

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

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

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

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

15. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention web designers: You have 50

milliseconds to make a good first impressoin! *Behavior & information technology*, 25, 2 (2006), 115-126.

16. Ling, C., Hwang, W., and Salvendy, G. A survey of what customers want in a cell phone design. *Behaviour & Information Technology*, 26, 2 (March-April 2007), 149-163.

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

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

19. McCay-Peet, L., Lalmas, M., and Navalpakkam, V. On saliency, affect and focused attention. In *SIGCHI Conference on Human Factors in Computing Systems* (2012), ACM, 541-550.

20. McDougall, S. and Reppa, I. Ease of icon processing can predict icon appeal. In *15th International Conference, HCI International* (2013), Springer Berlin Heidelberg, 575-584.

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

22. Miniukovich, A, and De Angeli, A. Computation of Interface Aesthetics. In *CHI'15* (Seoul 2015).

23. Miniukovich, A. and De Angeli, A. Quantification of Interface Visual Complexity. In *the 2014 International Working Conference on Advanced Visual Interfaces* (Como 2014), ACM, 153-160.

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

25. Miniukovich, A. and De Angeli, A. Visual Impression of Mobile App Interfaces. In *NordiCHI'14* (2014), ACM.

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

27. Mudambi, S. M. and Schuff, D. What makes a helpful review? A study of customer reviews on Amazon.com. *MIS Quarterly*, 34, 1 (2010), 185-200.

28. Nadkarni, S. and Gupta, R. A task-based model of perceived website complexity. *Mis Quarterly* (2007), 501-524.

29. Navalpakkam, V. and Itti, L. Modeling the influence of task on attention. *Vision Research*, 45, 2 (2005), 205-231.

30. Ng, A. W. and Chan, A. H. Visual and cognitive features on icon effectiveness. In *international multiconference of engineers and computer scientists* (2008), 19-21.

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

32. Parkhurst, D. J. and Niebur, E. Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, 19, 3 (2004), 783-789.

33. Purchase, H. C., Freeman, E., and Hamer, J. An exploration of visual complexity. *Digrammatic representation and inferences* (2012), 200-213.

34. Reber, R., Schwarz, N., and Winkielman, P. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8, 4 (2004), 364-382.

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

36. Reppa, I. and McDougall, S. When the going gets tough the beautiful get going: aesthetic appeal facilitates task performance. *Psychonomic bulletin & review*, 22 (2015), 1243-1254.

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

38. Sonderegger, A., Zbinden, G., Uebelbacher, A., and Sauer, J. The influence of product aesthetics and usability over the course of time: a longitudinal field experiment. *Ergonomics*, 55 (2012), 713-730.

39. Tamura, H., Mori, S., and Yamawaki, T. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8, 6 (1978), 460-473.

40. Thielsch, M. T. and Hirschfeld, G. High and low spatial frequencies in website evaluations. *Ergonomics*, 53, 8 (2010), 972-978.

41. Thielsch, M. T. and Hirschfeld, G. Spatial frequencies in aesthetic website evaluations– explaining how ultra-rapid evaluations are formed.

Ergonomics, 55, 7 (2012), 731-742.

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

43. Tinio, P. P. and Leder, H. Just how stable are stable aesthetic features? Symmetry, complexity, and the jaws of massive familiarization. *Acta Psychologica*, 130, 3 (2009), 241-250.

44. Tractinsky, N. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In *the ACM SIGCHI Conference on Human factors in computing systems* (1997), ACM, 115-122.

45. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the consistency of immediate aesthetic perceptions of web pages. *International journal of human-computer studies*, 64, 11 (2006), 1071-1083.

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

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

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

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

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

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