Representing visual aesthetic tastes for web pages in computable form

Lahiru G. Jayatilaka
Stanford University
lahiru@stanford.edu

Advisors: Dave T. Jackson, Scott R. Klemmer, Sebastian Thrun

Author Keywords

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces

General Terms
Human Factors

ABSTRACT
Research suggests that the visual attractiveness of a website is a strong determinant of users' online behavior. However, designing visually attractive websites is difficult. Among other challenges, web designers have to balance aesthetic design decisions against usability objectives, and account for the increasingly diverse range of display platforms (smartphones, tablets, etc.) on which web pages can now be viewed on.

Intelligent design support systems are a potential solution to this problem. However, to intelligently assist designers in creating attractive web pages, computer systems need to possess *computable representations* of human visual aesthetic tastes. Using machine-extracted style-based features and visual aesthetic judgements from 18 volunteers on 56 websites, we present preliminary results that suggest that such representation may be possible.

INTRODUCTION AND MOTIVATION
Recent research suggests that the visual aesthetics of web interfaces is a strong determinant of users' online behavior. In a study assessing trustworthiness of online health information sites, visually unappealing sites were rejected within a few seconds, whereas more appealing sites were scrutinized for content before being either accepted or rejected [12]. In the context of e-commerce, research has also indicated that the visual aesthetic quality of websites impacts intention of consumers to purchase products [1, 2]. The cited studies suggest that visual design choices such as low color contrast between text and background, large amount of text on a page and poor use of color are some sources for variation in user behavior.

Designing aesthetically pleasing web sites is challenging for several reasons. Cultural differences in aesthetic tastes pose a significant—possibly insurmountable—challenge to designers aiming to appeal to broad audiences [9]. Furthermore, designers must also balance aesthetic design decisions against other design objectives. For example, designers must make decisions about the amount of text on a page versus the (popular) aesthetic preference for "clean designs", or decisions about the load time of a web page against the quality of graphics to include on the page. The diversity of computing platforms has also increased the need for websites that "*look good*" not only on a laptop screen but also on a 3.5" smartphone screen and a 10" tablet screen, making the job of a web designer even more challenging.

A potential solution to these challenges are intelligent design support tools that can assist in creating visually attractive designs. These tools should be able to provide designers with creative inspiration through presentation of appropriate examples [10], directly suggest aesthetics-based improvements such as "reduce the amount of text," "increase the color contrast between background and text," "reduce the number of images" etc., to automatically generating more visually attractive alternatives of the current design. However, a prerequisite for such systems is the capacity to represent *visual aesthetic tastes in a computable form*. More specifically, in order to offer the envisioned level of design support, computer systems should be able to represent human visual aesthetic tastes in terms of variables in the design space. Using visual aesthetic judgements from 18 volunteers on 56 web designs, along with statistical learning techniques, we demonstrate that such representation may be feasilbe.

In the remainder of this paper we summarize prior work in the area of statistical learning of "visual appeal" for web designs. We then describe an experiment to investigate the possibility of representing visual aesthetic tastes for web designs in a computable form. We conclude with a discussion of our findings and recommendations for future work in this area.

.

Use of Space	DOM tree depth (min, max, mean, stddev), # of DOM leaf nodes, document width/height, amount of separation between content blocks, foreground/background ratio, overlapping element area, connected components in rendered image (#, minarea, maxarea, meanarea)
Use of Color	Color (mean, stddev), saturation (min, max, mean, stddev), value (min, max, mean, stddev), # of colors (in DOM, in rendered page), most dominant color, most dominant text color, # of dominant colors, text-to-background contrast (min, max, mean, stddev), histogram
Use of Text	# of words in page, # of words per block (min, max, mean, stddev), # of fonts, font size (min, max, mode, mean)
Use of Images	# of images, aspectRatio (min, max, mean, stddev), area (min, max, mean, stddev), complexity (min, max, mean, stddev)

Figure 1. Our statistical models to predict visual attractiveness of web pages were built using style-based features. These style-based features describe a designs use of space, color, text, and images, and are informed by interviews with designers and principles from the design literature, (See d.Tour (Ritchie 2011) for more details)

PREVIOUS WORK

Previous research has demonstrated that rapid aesthetic judgments of a website such as *repelling vs appealing* and *complicated vs simple* correlate well with specifically quantized low-level image-based features such as *color, texture, and intensity* of a web page [16]. The authors of the aforementioned work suggest that it is possible to develop computer algorithms that can deliver "quick and dirty evaluations' of subjective aspects of any (image-based) design. In a different context, the same researchers also demonstrated that a supervised learning algorithm was capable of learning the correlation between low-level image statistics, such as spatial frequency, luminance entropy, etc., and users' judgments of perceived usability and subjective appearance of car infotainment systems [15]. However, neither of these papers conclusively prove that visual aesthetic tastes for web pages can be represented in computable form.

We extend this previous research in two novel and important ways. First, we learn statistical models for predicting visual aesthetic tastes for web pages in terms of *style-based* features which have already been employed in engineering intelligent design support tools [3, 10]. These *style-based* features describe a design's use of space, color, text, and images, and are informed by interviews with designers and principles from the design literature [10] (Figure 1). Second, we systematically demonstrate through careful performance evaluation that it may be possible for style-based feature models to learn—represent in a computable form—visual aesthetic tastes.

EXPERIMENT

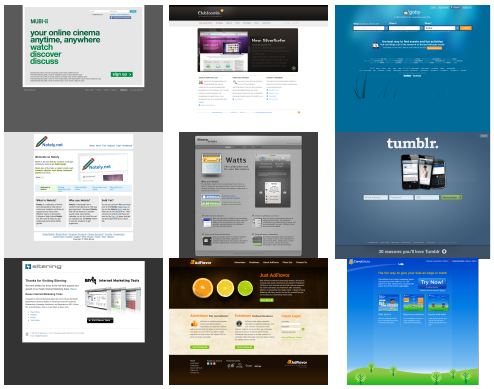
We conducted an experiment to investigate whether *style-based* feature (Figure 1) models can reliably learn and predict visual aesthetic tastes on 54 websites from two domains of web page designs (*Design domain* and *Software domain*). Specifically, we formulated the following two hypotheses:

H.1 *Style-based* feature models can reliably learn visual aesthetic tastes for a group of people in the *Design domain*

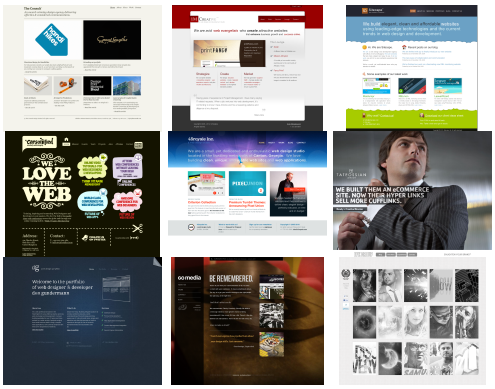
H.2 *Style-based* feature models that are learned on the *Design domain* can make reliable predictions on the *Software domain*

In the next sections we describe an experiment to collect visual aesthetic judgments on the experimental dataset, our approach to estimating *group-level* visual aesthetic tastes and results from building and evaluating three supervised statistical models.

Note on terminology: Visual aesthetic judgments will refer to the aesthetic opinions collected from our volunteers (or expressed at group-level), and visual aesthetic tastes will refer to the phenomena that we are trying to learn and represent in a computable form



(a) Select designs from Software domain



(b) Select designs from Design domain

Figure 2. Select designs used for experiment from the *Software domain* **2(a) and from the** *Design domain* **2(b)**

Collecting visual appeal judgements

Data set of web pages

Fifty-four web pages from two domains were ranked and rated in this study. 20 of these pages were of companies offering downloadable or internet-based software packages (Figure 2(a)),while 34 of these pages were of companies or individuals offering web design services (Figure 2(b)). We

Figure 3. A participant rank-ordering designs from the *Design domain*

used the web page corpus from [10], where all pages were selected from design blogs and award lists . Using the Qt framework in C++, each page was screen captured based on its rendering on a web browser at a resolution of approximately 1280 x 1024 pixels[1]. The designs were printed using a 300 DPI color laser printer on $8\frac{1}{2}^{"}$ x $11^{"}$ laser printer paper.

Volunteers

Eighteen full-time U.S. graduate students volunteered for this experiment. They included 16 men and 2 women aged between 21 and 28. All participants had normal or corrected-to-normal vision, and were free of color blindness. All participants had limited background in web design. Participants were offered candy as compensation for their participation.

Procedure

In order to measure visual aesthetic judgements on the corpus of web designs we designed a two phase paper-based *rank-ordering experiment*. Before the experiment, volunteers were informed that they would have to physically order, according to "visual attractiveness", 20 designs for *'websites marketing web-based or downloadable software packages'*(*Software domain*), and then repeat the same process for 34 designs from *'websites offering web-design services'* (*Design domain*). Volunteers were also informed that they would be working under time pressure. This experimental design choice was made in order to prevent volunteers' decisions being by swayed text and other informational content on a page and was informed by the related experiments measuring visual aesthetic judgements for web pages [5, 6, 4, 16]. Time limits were set for this experiment after running 4 pilot subjects. To motivate our volunteers we primed them with the following prompt:

You've been an avid web user for the past ten years (at least). Over this period you've developed a keen sense of what makes a website look good. Consider this experiment a platform for sharing some of your intuition with us

[1]Resolutions of captured sites varied marginally depending on how the web page was rendered

In the first phase, under a time limit of 4 minutes, participants rank ordered the 20 designs from the *software domain* on the floor according to "visual attractiveness" (participants were free to decide on the spatial arrangement of their orderings) (Figure 3). Time warnings were provided with 120, 60, 30, 20, 10, and 5 seconds remaining. Phase 2 was identical to Phase 1 but with 34 designs from the *design domain* and time-limit of 6 minutes. Time warnings were provided with 240, 120, 60, 30, 20, 10, and 5 seconds remaining.

Analysis

1. Clustering visual appeal judgements at group-level

We applied K-means on the *median* and *standard deviation of rank position* of each design to generate clusters of '*appealing*' and '*unappealing*' designs at the *group-level*. We considered the median rank of a design as representing its *degree of visual aesthetic appeal* at the group-level, and its standard deviation as representing the *level of agreement* within the group about the designs visual aesthetic appeal. Conceptually, a design with low median rank and low standard deviation is considered an '*appealing* design with *high group-level agreement'* , while a design with high median rank and relatively large standard deviation is considered an '*unappealing* design with *low group-level agreement*'.

2. Model building and statistical testing

After generating *group-level* (binary) labels, we trained and tested three types of supervised learning models: Naive Bayes [8], Binary logistic regression [7] and Linear Kernel Support Vector machines [11]. To test our experimental hypothesis (See Page 2) we evaluated each model using two measures of performance:

a. Within-domain performance (H1) : Leave One Out Cross Validation error on *Design domain*

b. Transfer performance (H2): Test error on *Software domain* of model trained on *Design domain*

Due to the small number of examples in both domains, we performed statistical confidence testing on each model's performance measures. More specifically, we considered a model to have learned *group-level* visual aesthetic tastes with statistical significance performance if and only if the following null hypothesis could be rejected using a 95% confidence interval ($p<0.05$):

H_0 : Learned model error $\geq$ Coin-flipping model error

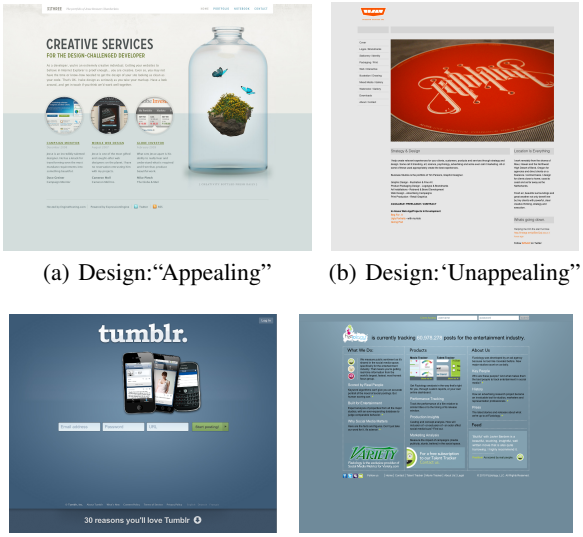
Coin-flipping model error is the expected performance of a model that makes decisions by randomly flipping a coin.

We used the 'exact test for goodness-of-fit' [13] to test for statistical significance.

Results

Group-level 'appealing' and 'unappealing' clustering

K-means with squared euclidian distance was run until convergence. Clusterings in both domains were linearly separable. (Figure 4 shows some results of K-means).



(a) Design:"Appealing" (b) Design:'Unappealing"

(c) Software:'Appealing" (d) Software:"Unappealing'

Figure 4. Results of k-means clustering approach on *Design* and *Software domain*. 4(a) and 4(b) are designs that were identified by K-means as '*appealing*' and '*unappealing*' at a *group-level*, 4(c) and 4(d) are the analogous outcomes from the software domain.

In the design domain (n=34), the *'appealing'* cluster contained 17 designs
($\mu_{median\ rank,\ stddev\ rank} = (12.9, 8.6)$,
$\sigma_{median\ rank,\ stddev\ rank} = (3.60, 1.57)$)
while the *'unappealing'* cluster contained 17 designs
($\mu_{median\ rank,\ stddev\ rank} = (23.1, 9.3)$,
$\sigma_{median\ rank,\ stddev\ rank} = (3.30, 1.70)$)

In the software domain (n=20), the *'appealing'* cluster contained 7 designs ($\mu_{median\ rank,\ stddev\ rank} = (5.6, 4.5)$,
$\sigma_{median\ rank,\ stddev\ rank} = (2.50, 0.70)$),
while the *'unappealing'* cluster contained 13 designs
($\mu_{median\ rank,\ stddev\ rank} = (12.6, 5.0)$,
$\sigma_{median\ rank,\ stddev\ rank} = (2.00, 0.86)$)

Supervised model evaluation

The Naive Bayes model had a *within-domain performance* of 0.44 (False positive rate (FP) = 0.35, False negative rate (FN) = 0.53, not statistically significant (n.s.)) and *across-domain performance* of 0.55 (FP=0.29, FN=0.7, n.s.). The logistic regression model had a *within-domain performance* of 0.53 (FP = 0.59, FN = 0.47, n.s.) and *across-domain performance* of 0.65 (FP=0, FN=1.0, n.s.). The linear kernel SVM (KKT Tolerance =0.7, C=1) had a *within-domain performance* of 0.35 (FP= 0.35, FN = 0.35, p* < 0.05) and *across-domain performance* of 0.30 (FP= 0.29, FN = 0.30, p* < 0.05).

DISCUSSION

We observe that the standard deviation of the cluster groups are comparable along the dimensions of *degree of visual appeal* and *group-level agreement* for both domains. This observation suggests that K-means provided a reliable estimate of *group-level* aesthetic judgements, and furthermore that our analysis with binary labels was ecologically valid.

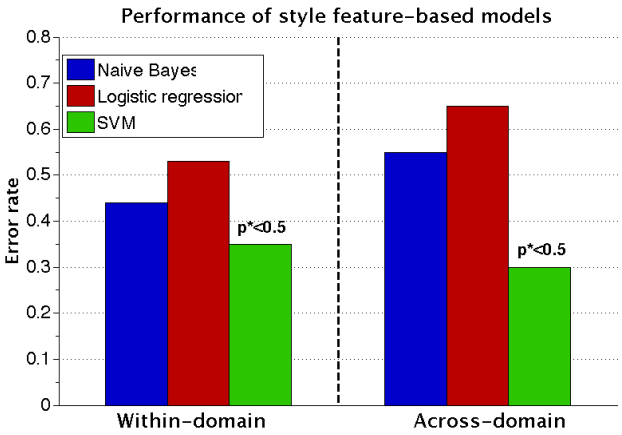


Figure 5. Graph shows performance of the three supervised learning models. The 'Within-domain' section shows the LOOCV performance on the *Design Domain*. The 'Across-domain' section shows the performance of the models trained on the *Design Domain* and tested on the *Software Domain*. The SVM model had statistically significant performance in both cases.

Our results show that a linear kernel support vector machine has statistically significant performance on both *within-* and *across- domain* measures. These results verify our experimental hypotheses **H1** and **H2** (See Page 2). Furthermore, the false positive and false negative rates also suggest that the model was finding statistical structure of *group-level* visual aesthetic tastes in terms of the *style-based features*. Further verification of learning may be pursued through feature selection on the learned SVM model [14] followed by qualitative analysis relating style-based design features to *group-level* visual aesthetic judgements.

A noticeable shortcoming of our model evaluation analysis is its lack of bi-directionality. We do not consider *within-* and *across-domain* measures from the perspective of models trained on the *Software domain*. This analysis choice was made given the relatively small number of examples (n=20) in the dataset for the domain.

CONCLUSION AND FUTURE WORK

We aimed to explore whether it was possible to express visual aesthetic tastes for web designs in a computable form. The findings from our experiment suggest that such representation may be possible. Given the limited scope of our investigation, further work is required to confirm the learnability of visual aesthetic tastes for web pages.

In order to more conclusively establish that visual aesthetic tastes for web pages can be expressed in a computable form, it is clear that future work in this area must address obvious issues of scale. We need to understand how style-based feature models perform with larger and more diverse populations, and also to investigate how these models transfer to more diverse domains (e.g. medical websites, news pages, discussion blogs, etc.). We hope that our work lays the foun-

dation for more extensive research in computational techniques that can be used to engineer intelligent design support tools.

ACKNOWLEDGEMENTS

I would like to thank Scott Klemmer and Sebastian Thrun who co-advised this work as a first year research rotation project. I would also like to thank Arvind Satyanarayan, Amrapali Maitra, Daniel Ritchie, Dave Jackson and Ranjitha Kumar and for their guidance and support.

REFERENCES

1. Everard, A., and Galletta, D. How presentation flaws affect perceived site quality, trust, and intention to purchase from an online store. *Journal of Management Information Systems 22*, 3 (2006), 56–95.
2. Hall, R., 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.
3. Kumar, R., Talton, J., Ahmad, S., and Klemmer, S. Bricolage: Example-based retargeting for web design. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, ACM (2011), 2197–2206.
4. Lavie, T., and Tractinsky, N. Assessing dimensions of perceived visual aesthetics of web sites. *International journal of human-computer studies 60*, 3 (2004), 269–298.
5. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Trans. Comput.-Hum. Interact. 18* (May 2011), 1:1–1:30.
6. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention web designers: You have 50 milliseconds to make a good first impression! *Behaviour & information technology 25*, 2 (2006), 115–126.
7. McCullagh, P., and Nelder, J. *Generalized linear models*. Chapman & Hall/CRC, 1989.
8. Mitchell, T. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* (1997).
9. Reinecke, K., and Bernstein, A. Tell me where youve lived, and ill tell you what you like: Adapting interfaces to cultural preferences. *User Modeling, Adaptation, and Personalization* (2009), 185–196.
10. Ritchie, D., Kejriwal, A., and Klemmer, S. d. tour: style-based exploration of design example galleries. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, ACM (2011), 165–174.
11. Shawe-Taylor, J., and Cristianini, N. An introduction to support vector machines and other kernel-based learning methods. *Cambridge University Press, UK* (2000).
12. Sillence, E., Briggs, P., Fishwick, L., and Harris, P. Trust and mistrust of online health sites. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2004), 663–670.
13. Sokal, R., and Rohlf, F. The principles and practice of statistics in biological research. *New York.: Edition 3* (1995).
14. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. Feature selection for svms. *Advances in neural information processing systems* (2001), 668–674.
15. Zheng, X., Chakraborty, I., Lin, J., and Rauschenberger, R. Developing quantitative metrics to predict users' perceptions of interface design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, SAGE Publications (2008), 2023–2027.
16. 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 *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, ACM (New York, NY, USA, 2009), 1–10.