Modeling How Different User Groups Perceive Webpage Aesthetics

Luis A. Leiva[1,*], Morteza Shiripour[2], Antti Oulasvirta[2]

[1] University of Luxembourg, Luxembourg
[2] Aalto University, Finland
[*] Corresponding author: name.surname@uni.lu

Abstract

Aesthetics is a central consideration in user interface design. It is known to affect end-user behavior and perception, in particular the first impression of graphical user interfaces. However, what users perceive as pleasant or good design is highly subjective. We contribute a computational model that estimates the visual appeal of a given webpage for several common cohorts, or user groups, including gender, age, and education level. Our model, a convolutional neural network trained on 418 webpage screenshots having 771k aesthetic scores (in a 1–9 Likert scale) from 32k users, achieves high accuracy and is always less than 1 point off from ground-truth ratings. Designers can use our model to anticipate how people would rate their webpage, offer personalized designs according to the visual preferences of their users, and support rapid evaluations of webpage design prototypes for specific cohorts.

Keywords: Webpage Aesthetics; Visual Design; Computational Accessibility; Neural Networks

1 Introduction

Aesthetics plays a key role for user behavior and perception, especially when it comes to forming a first impression of a given webpage [21, 22]. First impression refers to the perception of beauty or appeal formed within the first few seconds, and even milliseconds, within viewing a design [9]. It has been noted that aesthetics can impact user's decision making process [52] and consumer loyalty [39, 25] and it is found a strong determinant of users' satisfaction and pleasure [19]. Indeed, a pleasant design can attract more users, increase credibility, and influence purchase intentions [36, 46].

This paper contributes to understanding different users' variability in webpage aesthetics. We will refer to the aesthetics construct as *first impression of* *visual appeal*. What users perceive as pleasant or good design is highly subjective; people with different backgrounds usually have different visual preferences [38, 54]. For example, it has been reported that people in Russia and Finland prefer simpler designs, while people in North Macedonia and Malaysia prefer colourful designs [37]. Another study found that males tend to rank higher webpages designed by males, and similarly for females [31]. It is commonly agreed that our human visual system is able to quickly generate feature representations that can describe visual appeal [10]. At the same time, designers may have some target users in mind but eventually they evaluate their designs with few users, mainly because of limited time or budget. Therefore, first impression of web pages effectively creates a permanent impression [22]. It is hence important for designers to quantify webpage aesthetics reliably.

We contribute a computational model that estimates the visual appeal of a given webpage for several cohorts, i.e. user groups or profiles. To the best of our knowledge, there is no previous model to quantify first impression *conditioned* to particular characteristics of a cohort, for example gender, age, or education level. Previous work assumed a one-size-fits-all homogeneous user group, however people with different backgrounds have different perceptions of design aesthetics. Reinecke et al. [38] investigated the statistical effect of demographic information on visual appeal, but they never predicted the perceived aesthetics score for any user cohort. Our model, a convolutional neural network trained on 418 webpage screenshots having 771k aesthetic scores (in a 1–9 Likert scale) from 32k users, is highly accurate and always less than 1 point off from ground-truth scores. Ultimately, our model provides data-driven insights which designers can use to judge and compare webpage designs, anticipate how people would rate their webpage, and support rapid evaluations of webpage design prototypes for specific cohorts.

2 Related Work

The research literature on visual aesthetics is huge and space precludes a complete treatment. We would recommend the review by Tractinsky et al. [47] and a more recent meta-analysis by Thielsch et al. [45]. Owing to the importance of first impression, several general guidelines have been proposed for designing a webpage [1, 35, 40]. However, it is also known that there is vast individual variability in aesthetics preferences [21, 22]. Previous work has attributed variability to demographic attributes and culture [37], personal traits [14], as well as gender [31]. To avoid fixating on a single user group, it is important that designers can take such variability into account. This calls for methods that help designers flexibly take different stances beyond the immediately available ones.

However, how to best quantify aesthetics is an ongoing research topic. In bottom-up approaches, aesthetics is computed using image features. Researchers have mainly focused on two hand-crafted visual features: colorfulness [6, 29] and visual complexity [27, 51, 56]. However, there are more visual features that may impact the user's appealing perception, such as symmetry [50], compositional elements [2], wireframe geometry [26], balance [20], harmony [30], novelty [12], and typography [23]. Quantifying all these features is hard because there is no measurement consensus; see e.g. Miniukovich et al. [28].

In top-down approaches, human visual appeal is used to train a statistical or machine learning model. However, classic statistical models (see e.g. [38, 55]) are often unable to model complex relationships and capture non-linearities within the data, especially if we want to condition the model to different cohorts. Hence, a more reasonable approach is to predict self-reported measures of webpage aesthetics with deep learning models, considering the tremendous impact they have had in computer vision and related areas [53]. Concretely, convolutional neural networks (CNNs) currently set the state of the art in visual recognition [3, 11] and have been proposed for computing webpage aesthetics before us [8, 18], however no previous work have considered any cohort types. As stated in the previous section, it is important for designers to quantify webpage aesthetics reliably, as different people perceive different designs in a very different way. In the following section we describe our approach.

3 Experiments

We use a public dataset from LabintheWild[1] that contains ground-truth ratings (1–9 Likert scale) of first-impression visual appeal for 418 webpage screenshots from about 32k participants [37]. The dataset provides 771083 paired ratings together with demographic information such as gender, age, education level, language, and citizenship.

As explained by Reinecke et al. [37], screenshots were displayed for 500 ms and were rated twice by each participant. Thus, we average both ratings so that each participant is considered as an independent rater of each webpage. Then, for each webpage we average all scores received from each participant, since the task is webpage's aesthetics prediction (not user's). Aesthetic scores are thus continuous values distributed in the $[1, 9]$ range,[2] so we framed our prediction task as a regression problem.

3.1 Model Architecture

To begin, we investigate what is the most adequate neural network architecture for our task, for which we train a general-purpose CNN model. In a nutshell, CNNs use a hierarchy of layers that progressively extract abstract visual features (feature maps) such as contours, borders, shapes, or textures, and propagates those features to subsequent layers, similar to the human vision system [16, 49]. We experiment with several deep learning models pre-trained on the popular ImageNet dataset [7], which we fine-tune to the LabintheWild dataset via transfer

[1] `http://labinthewild.org/studies/aesthetics/`

[2] Once aggregated by webpage, scores are roughly comprised in the $[1, 7]$ range.

learning [33, 43]. These pre-trained models are the following, by chronological order of discovery:

Inception [42], also know as GoogLeNet, is one of the earliest CNN architectures. It uses adaptive filters inspired by the Hebbian theory from human learning [24].

OxfordNet [41], also known as VGG-16, promotes small receptive fields and a simple, homogeneous architecture. It is still considered to be an excellent human vision model.

ResNet50 [13], named after its 50 layers stack, introduced residual learning via shortcut connections, allowing for faster training times and increased accuracy than its predecessors.

DenseNet [15] concatenates previous information from earlier convolutional layers and passes its feature maps to all subsequent layers. It achieved comparable accuracy to ResNet50 using fewer trainable weights (26M vs 32M, respectively).

Xception [5] uses depth-wise separable convolutions, so the number of connections are fewer and the model is lighter. It outperformed all its predecessors (including OxfordNet, ResNet50, and Inception) in several image classification tasks.

We also train our own CNN model, inspired by the OxfordNet architecture. Our custom CNN model has 2 convolutional layers with 32 filters of size 5, followed by a global average pooling layer, for regularization purposes, a fully-connected (FC) layer of 4096 neurons with 0.5 dropout, and the output layer with one neuron (since the aesthetics score is a single value). We tried other model configurations but they did not perform so well. For example, more than 2 layers did not improve performance and in some cases led to overfitting issues, which we resolved with the Dropout layer. The most important design decision is the receptive field of our CNN layers (size 5). By having such small receptive fields, it is possible for different hidden neurons to become highly specialized in specific regions of the input image. We tried more filters and larger filter sizes, but they did not improve performance either. All layers use ReLU activation [32], except the output layer which uses linear activation (since it is a regression model). Finally, we also consider the Webthetics model proposed by Dou et al. [8], which is the closest approach to our work and whose implementation is publicly available.[3]

3.2 Model Training

We randomly split the 418 screenshots in three partitions: a training set with 267 screenshots, a validation set with 67 screenshots, and a test set with the

[3] `https://github.com/carrenD/Webthetics`

remaining 84 screenshots. We use stratified sampling to ensure our partitions are well balanced (Figure 1). We train all models (Table 1) on the training set for 200 epochs at most, using early stopping of 10 epochs to retain the best model weights and monitor its performance on the validation set. We do not apply any data augmentation technique (e.g. cropping or horizontal flipping) since we argue that webpage aesthetics should be assessed according to the original webpage design and not a modified version of it. We train in batches of 32 screenshots each, which are resized to a 160x120 px resolution to speed up training. We use the popular Adam optimizer with learning rate $\eta = .0005$ and decay rates $\beta_1 = 0.9$ $\beta_2 = 0.99$, and the mean squared error (MSE) as a loss function. Finally, the trained models are evaluated on the held-out test set, which simulates unseen data.

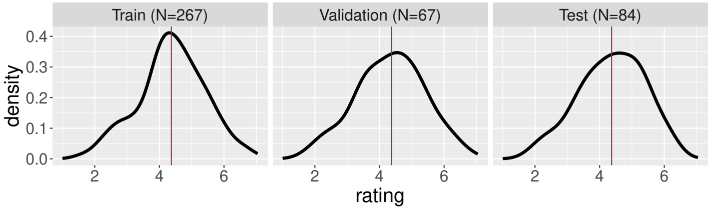


Figure 1: Distribution of user ratings in each data partition. The vertical red line denotes the mean rating.

4 Results

The overall results of these experiments are reported in Table 1. Together with the MSE, we also report the mean absolute error (MAE). While MSE penalizes large deviations in the model predictions, MAE informs about the overall prediction variability. We also discretize the aesthetic scores by rounding to the nearest integer, to simulate what a user would have rated in the original LabintheWild setting, and compute the multi-class accuracy score. A random classifier would be right in 11.1% of the cases, therefore any competitive model should be more accurate than this.

As can be observed in Table 1, both Webthetics and our custom CNN outperform all the pre-trained models. We can see that all pre-trained deep learning models are off by more that one point in terms of MSE and more than 2 points in MAE, therefore when rounding the predicted aesthetic scores to the nearest integer, they do not match the ground-truth scores in most cases, and so accuracy is 0.0 for some of these models. We attribute the bad performance of such pre-trained models to the fact that they were exposed to natural scene imagery [7], and webpage screenshots are of quite a different nature; e.g. they are more colorful, have sharper borders, often less textures, etc. Hence, by training a CNN model from scratch (both Webthetics and our custom model),

Table 1: Experiments to decide the most adequate network architecture. As a reference, a random classifier would achieve $1/9 = 11.1\%$ accuracy on this task.

		Test set performance		
Architecture	**Ref.**	**MSE** ↓	**MAE** ↓	**Acc. (%)** ↑
Inception	[42]	1.143	2.740	6.2
OxfordNet	[41]	1.289	2.945	3.1
ResNet50	[13]	1.376	3.118	0.0
DenseNet	[15]	1.376	3.118	0.0
Xception	[5]	1.396	3.097	0.0
Webthetics	[8]	0.749	0.704	45.2
Our CNN model		**0.678**	**0.630**	**57.1**

it is possible to learn these subtle but important design patterns. In addition, pre-trained models are far more complex than our custom CNN and thus require much more training data to achieve competitive performance, even after fine-tuning (transfer learning) to the task at hand.

It is worth mentioning that our custom CNN model outperforms Webthetics by a small margin in terms of MSE and MAE, however the gain in terms of classification accuracy is clear, with more than 10 percentual points of difference. More importantly, our custom model architecture is much simpler: we only use two CNN layers and one FC layer, 167k trainable weights in total, whereas Webthetics uses five CNN layers (with max. pooling and dropout) and two FC layers, 5M weights in total. We also note that the original Webthetics paper [8] used 256x192 grayscale webpage screenshots as input and was pre-trained on 80k Flickr images. Further, Webthetics did not use a held-out test partition (which simulates unseen data) so, taken together, these differences explain the slightly worse performance in our experiments as compared to their paper. In sum, our model outperforms state-of-the-art approaches in this regression task with a significant reduction in computational complexity (30x less model weights).

The MAE of our custom CNN model is 0.63, which indicates that the predictions delivered by the model are off by less than one aesthetic score point on average. In other words, if a given webpage has received an aesthetics score of, say, 6 on average, our model most probably will predict a score that is greater than 5.37 and less than 6.63. Then, considering that $\sqrt{\text{MSE}} = 0.82$, we can see there is some variation in the magnitude of the errors though very large errors are unlikely to have occurred; e.g. because of outliers. Finally, the Spearman's rank correlation coefficient is $\rho = 0.637$ ($p < .001$), which indicates that our model's predictions correlate well with ground-truth ratings.

4.1 Cohort-based Results

We are now confident that our model can predict with notable accuracy the visual appeal score of a given webpage for a "generic user" group. But what about predicting for different user cohorts? We repeat the training procedure of our custom CNN model while segregating the data on the basis of the following demographic information available: gender, age, and education categories. Table 2 depicts the results of these experiments.

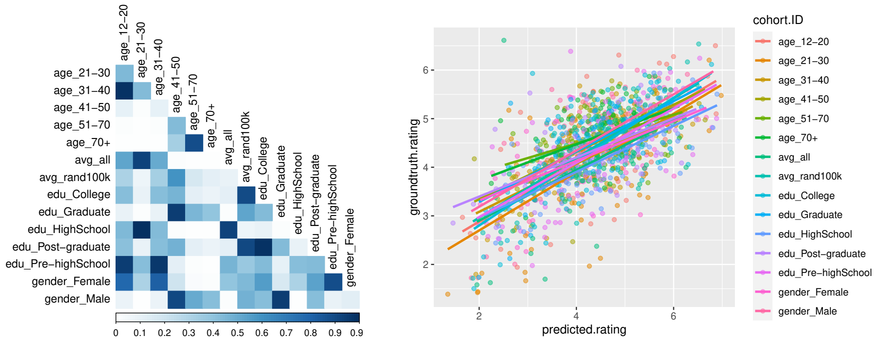
We can see that our model is still very accurate, with a MAE below 0.7 and an MSE no larger than 0.163, which corresponds to a deviation of 0.4 points on average. Again, we can conclude that our model's predictions correlate well with ground-truth ratings, as further indicated by the Spearman's rank correlation coefficient. All correlations reported in Table 2 are statistically significant ($p < .001$).

Table 2: Test performance results of our model for all the different cohorts considered. We denote in bold typeface the best result column-wise within a given cohort.

		Num Ratings			**Test performance**		
Category	**Cohort**	**Train**	**Validation**	**Test**	**MSE** $\downarrow$	**MAE** $\downarrow$	$\rho \uparrow$
Gender	Male	221044	55261	69076	**0.114**	0.662	**0.630**
	Female	267300	66825	83532	0.129	**0.644**	0.553
Age	12–20	73013	18254	22817	0.122	0.695	0.658
	21–30	190736	47684	59605	0.163	0.778	0.562
	31–40	110065	27517	34396	0.113	0.638	0.610
	41–50	54178	13545	16931	**0.089**	**0.513**	**0.700**
	51–70	51928	12982	16228	0.109	0.579	0.402
	+70	3494	874	1092	0.119	0.587	0.466
Education	Pre-high school	6142	1536	9598	0.123	0.661	0.637
	High school	71006	17752	22190	0.128	0.634	**0.724**
	College	205179	51295	64119	0.103	**0.576**	0.641
	Graduate school	118740	29685	37106	**0.100**	0.594	0.598
	Post-graduate	45614	11404	14255	0.142	0.691	0.539
All	General	493494	123373	154216	0.678	0.630	0.637
Rand. sample	100k	64000	16000	20000	0.088	0.556	0.665

Age-related changes in visual perception are well known in the research literature, see e.g. [1, 51, 29, 6]. Interestingly, while Spearman correlation in our '+70' cohort is smaller than in other cohorts, the MAE is sensibly smaller. Overall, we can see that people aged 40 or older exhibit less variance in their perceived aesthetics scores. No sensible differences in terms of MSE were observed across all the cohorts considered.

We performed a statistical analysis to see if there is some difference in the predictions delivered for any of the user cohorts. We include two baselines in these comparisons: the average model (trained on all the available data) and a model trained on 100k random data points. First, we run an analysis of variance (ANOVA) as omnibus test: $F(14, 1162) = 4.03, p < .001$. Since the omnibus test is significant, we run pairwise t-tests as post-hoc tests, to see where exactly are the differences between conditions. We apply the false discovery rate (FDR)



(a) Pairwise comparisons heatmap

(b) Aesthetics score prediction

Figure 2: Left: p-values (FDR-corrected) of the pairwise comparisons after a significant omnibus test. Lighter colors denote smaller p-values. Right: predicted vs. ground-truth aesthetic scores in the test set.

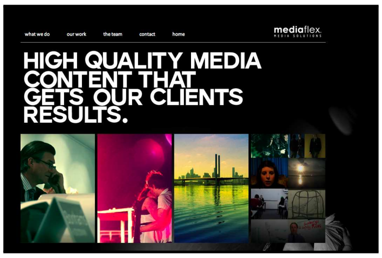
correction to guard against type I errors because of multiple comparisons. FDR has more statistical power than family-wise correction methods like Bonferroni. Figure 2(a) summarizes the results. As can be observed, differences between groups are statistically significant in many cases. In sum, having group-specific models instead of having just a single (general) model is beneficial and definitely more informative that a one-size-fits-all approach.

4.2 Prediction examples

Finally, we provide prediction examples for webpages that were perceived as good and bad (Figure 3) from the test set, which means they were not seen during model training. As can be observed, our model achieves notably good performance for all cohorts, very similar to the performance achieved by the general model and mostly in line with ground-truth aesthetic scores. This is also reflected in Figure 2(b), where we can see that all trend lines are often aligned with the diagonal of the plot. We can conclude that our model can predict first-impression webpage aesthetics for different user cohorts reliably. Although the differences in how people rate website aesthetics may seem small at first sight, we argue that it is interesting to see how different cohorts differ from one another, even if there is only one or two aesthetics score points of difference. More important, our model can provide an estimation range for a given category (e.g. age), which is much more informative than providing a single prediction value for everybody.

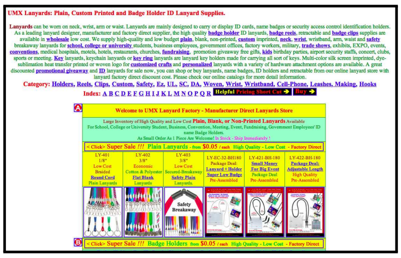
5 Discussion and Future Work

In this paper, we have studied a data-driven method that can help take the perspective of another user group in design. Instead of automating design, we



HIGH QUALITY MEDIA CONTENT THAT GETS OUR CLIENTS RESULTS.

Cohort	True	Predicted
General	7.04	6.62
Male	7.19	6.68
Female	6.84	5.99
12–20 yr.	7.58	7.27
21–30 yr.	7.26	6.86
31–40 yr.	6.90	5.97
41–50 yr.	6.66	5.90
51–70 yr.	6.28	5.99
+70 yr.	5.74	5.67
Pre-high school	7.58	7.25
High school	6.97	6.54
College	7.06	6.90
Graduate school	7.18	6.67
Post-graduate	6.72	6.48



UMX Lanyards: Plain, Custom Printed and Badge Holder ID Lanyard Supplies.

Lanyards can be worn on neck, wrist, arm or waist. Lanyards are mainly designed to carry or display ID cards, name badges or security access control identification holders. As a leading lanyard designer, manufacturer and factory direct supplier, the high quality badge holder ID lanyards, badge reels, retractable and badge clips supplies are available in wholesale low cost. We supply high-quality and low budget plain, blank, non-printed, custom imprinted, neck, wrist, wristband, arm, waist and safety breakaway lanyards for school, college or university students, business employees, government offices, factory workers, military, trade shows, exhibits, EXPO, events, conventions, medical hospitals, motels, hotels, restaurants, churches, fundraising, promotion giveaway free gifts, kids birthday parties, airport security staffs, concert, clubs, sports or meeting. Key lanyards, keychain lanyards or key ring lanyards are lanyard key holders made for carrying all sort of keys. Multi-color silk screen imprinted, dye-sublimation heat transfer printed or woven logo for customized crafts and personalized lanyards with a variety of hardware attachment options are available. A great discounted promotional giveaway and ID lanyards for sale now, you can shop or buy lanyards, name badges, ID holders and retractable from our online lanyard store with lanyard factory direct discount cost. Please check our online catalogs for more detail information.

Category: Holders, Reels, Clips, Custom, Safety, Ez, UL, SC, DA, Woven, Wrist, Wristband, Cell-Phone, Leashes, Making, Hooks
Index: A B C D E F G H I J K L M N O P Q R
Helpful Pricing Short Cut
Buy
A

Welcome to UMX Lanyard Factory - Manufacturer Direct Lanyards Store

Large Inventory of High Quality and Low Cost Plain, Blank, or Non-Printed Lanyards Available
For School, College or University Student, Business, Convention, Meeting, Event, Fundraising, Government Employees' ID name Badge Holders.
As Small Order As 1 Piece Are Welcome! In Stock - Ship Immediately !

< Click> Super Sale !!! Plain Lanyards - from $0.05 / each High Quality - Low Cost - Factory Direct

LY-401 1/8" Low Cost Braided Round Cord Plain Lanyards	LY-402 3/8" Economic Cotton & Polyester Flat Blank Lanyards	LY-403 3/8" Low Cost Secured-Breakaway Safety Plain Lanyards.	LY-EC-32-BH180 Package Deal: Lanyard + Holder Super Low Budge Pre-Assembled	LY-421-BH-180 Small Money For Big Event Package Deal Pre-Assembled	LY-422-BH-180 Package Deal: Adjustable Length High Quality Pre-Assembled

< Click> Super Sale !!! Badge Holders from $0.05 / each High Quality - Low Cost - Factory Direct

Cohort	True	Predicted
General	1.48	1.83
Male	1.49	1.96
Female	1.47	1.94
12–20 yr.	1.39	1.99
21–30 yr.	1.35	1.38
31–40 yr.	1.43	2.00
41–50 yr.	1.54	2.11
51–70 yr.	2.05	3.50
+70 yr.	3.03	3.50
Pre-high school	1.84	2.48
High school	1.71	1.81
College	1.40	1.60
Graduate school	1.47	1.75
Post-graduate	1.50	1.88

Figure 3: Sample average ground-truth ratings (True column) vs. model predictions (Predicted column) for webpage designs perceived as *visually good* (right) and *visually bad* (left), picked at random from our test set.

believe it is necessary to study machine learning methods that are plausible, compatible, and can support designers' work. Such approaches in general may help in being more emphatic, but also in systematically evaluating alternative perspectives, so as to avoid fixating on a single user group or scenario.

We have addressed the challenging task of predicting webpage aesthetics (first impression of visual appeal) for different user cohorts. With our computational model, designers can now quickly get an analytical estimation of how different user groups would perceive a webpage. This can help designers to create more inclusive and more accessible designs (e.g. better layouts and/or color schemes for aging groups) informed by the quantitative predictions delivered by our model. The model can be integrated into design tools that permit plugins or APIs, such as Sketch[4] or Figma.[5]

[4] https://www.sketch.com/

[5] https://www.figma.com/

It is important not to confuse the presentation time of a given stimuli and the needed processing time of the human visual system. Previous works have presented stimuli for only 50 ms or even less [22, 44, 48, 51] and evaluations of these shortly presented stimuli were quite stable, but it is unlikely that the cognitive processing of these stimuli only takes a few milliseconds [34]. With respect to very quickly made aesthetic webpage evaluations, Bolte et al. [4] noted that it takes several hundred milliseconds to form first impressions, about 600–800 ms; which is not very far from the 500 ms onset used in the dataset we have analyzed in this paper.

One limitation of our model is that it cannot predict how an *individual* user would perceive a given webpage design, since our training data comprise aggregated scores from several users. We believe that individual-level prediction is presently out of reach, mainly due to unavailability of suitable training data. Training individual-level models would only be possible if we could collect several ratings from the same user for a large number of webpages.

Our proposed model architecture is rather simple, for which the LabintheWild aesthetics dataset is an adequate resource. Importantly, while the number of websites in this dataset can be deemed as small (418 screenshots), the number of aesthetics scores is very large: 771k scores coming from 32k users worldwide. Thanks to this large pool of user ratings, our model can accurately predict how different user cohorts would rate a given website. However, since the LabintheWild aesthetics dataset was released in 2014, it remains unclear how well those predictions would transfer to websites designed with modern CSS frameworks such as Material Design[6]. We believe that, if the rendering style is not dramatically different from those found in the LabintheWild dataset, our model should deliver on-par high-quality predictions for newer websites.

As noted, we did not investigate new design features to be used in computation, as these were automatically derived by our model. This is perhaps the most interesting property of deep learning models: in the past, a huge investment on manually crafted featured was required to train classic machine learning models. This is no longer the case with deep learning, where models can learn the best feature representation for the task at hand. However, since web design is subject to trends and fashions, an interesting avenue for future work would be to analyze the evolving nature of aesthetic judgments over time. As noted, our model has been trained on a "static" set of websites, therefore we should collect more (and more recent) training data, not only in order to conduct such analysis, but also to keep our predictions updated.

At present, our CNN architecture can assess first-impressions for a limited set of individual cohorts. For future work, we plan to add more cohorts (e.g. citizenship) and combine them to predict more sophisticated outcomes; i.e. we would like to offer designers the possibility of guessing how e.g. "Uneducated males in their fourties" would rate a webpage design. Finally, aesthetics may change according to the user's familiarity, due to learning or changes in the task [17], but in this paper we only have considered the first impression of

[6] `https://material.io/develop/web`

the users, since it has been shown that users make lasting judgments about a website's appeal within a split second of seeing it for the first time [38].

6 Conclusion

We have contributed a computational model that estimates the visual appeal of a given webpage for several common cohorts, including gender, age, and education level. Our model, a convolutional neural network trained on 771k aesthetic scores (in a 1–9 Likert scale) from 32k users, achieves high accuracy and is always less than 1 point off from ground-truth ratings. Previously it was not possible to anticipate how people from different user cohorts would rate a webpage. In addition, web designers can use our model to offer personalized designs according to the visual preferences of their users, and support rapid evaluations of webpage design prototypes for specific cohorts.

Declarations

Funding

This work was supported by the Department of Communications and Networking – Aalto University, the Finnish Center for Artificial Intelligence (FCAI), the Academy of Finland through the Human Automata (grant 328813) and BAD (grant 318559) projects, the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001) and the European Innovation Council Pathfinder program (SYMBIOTIK project).

Competing interests

The authors have no competing interests to declare that are relevant to the content of this article.

Availability of data and materials

Our software is available at `https://luis.leiva.name/aesthetics/`.

Authors' contributions

L. A. Leiva: Conceptualization, Methodology, Software, Writing - Original Draft. **M. Shiripour**: Software, Validation, Writing - Original Draft. **A. Oulasvirta**: Writing - Reviewing and Editing.

References

[1] S. U. Ahmed, A. Al Mahmud, and K. Bergaust. Aesthetics in human-computer interaction: Views and reviews. In *Human-Computer Interaction. New Trends. LNCS 5610*. 2009.

[2] M. Bauerly and Y. Liu. Effects of symmetry and number of compositional elements on interface and design aesthetics. *Int. J. Hum. Comput. Int.*, 24(3), 2008.

[3] S. Bianco, L. Celona, P. Napoletano, and R. Schettini. Predicting image aesthetics with deep learning. In *Proc. ACIVS*, 2016.

[4] J. Bölte, T. M. Hösker, G. Hirschfeld, and M. T. Thielsch. Electrophysiological correlates of aesthetic processing of webpages: a comparison of experts and laypersons. *PeerJ*, 5, 2017.

[5] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. CVPR*, 2017.

[6] D. Cyr, M. Head, and H. Larios. Colour appeal in website design within and across cultures: A multi-method evaluation. *Int. J. Hum. Comput. Stud.*, 68(1–2), 2010.

[7] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[8] Q. Dou, X. S. Zheng, T. Sun, and P.-A. Heng. Webthetics: Quantifying webpage aesthetics with deep learning. *Int. J. Hum. Comput. Stud.*, 124, 2019.

[9] M. Douneva, R. Jaron, and M. T. Thielsch. Effects of different website designs on first impressions, aesthetic judgements and memory performance after short presentation. *Interact. Comput.*, 28(4), 2015.

[10] R. A. Epstein and C. I. Baker. Scene perception in the human brain. *Annu. Rev. Vision Sci.*, 5, 2019.

[11] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, L. Wang, G. Wang, J. Cai, and T. Chen. Recent advances in convolutional neural networks. *Pattern Recognit.*, 77, 2018.

[12] A. Haig and T. A. Whitfield. Predicting the aesthetic performance of web sites: What attracts people?, 2001.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.

[14] W. D. Hoyer and N. E. Stokburger-Sauer. The role of aesthetic taste in consumer behavior. *J. Acad. Mark. Sci.*, 40(1), 2012.

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. CVPR*, 2017.

[16] D. H. Hubel and T. N. Wiesel. Brain mechanisms of vision. *Scientific American*, 241(3), 1979.

[17] J. P. Jokinen, J. Silvennoinen, and T. Kujala. Relating experience goals with visual user interface design. *Interact. Comput.*, 30(5), 2018.

[18] M. G. Khani, M. R. Mazinani, M. Fayyaz, and M. Hoseini. A novel approach for website aesthetic evaluation based on convolutional neural networks. In *Proc. ICWR*, 2016.

[19] T. Lavie and N. Tractinsky. Assessing dimensions of perceived visual aesthetics of web sites. *Int. J. Hum. Comput. Stud.*, 60, 2004.

[20] D. Lawrence and S. Tavakol. *Balanced website design: optimising aesthetics, usability and purpose.* Springer Science & Business Media, Berlin/Heidelberg, Germany, 2006.

[21] G. Lindgaard, C. Dudek, D. Sen, L. Sumegi, and P. Noonan. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Trans. Comput. Hum. Interact.*, 18(1), 2011.

[22] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown. Attention web designers: You have 50 milliseconds to make a good first impression! *Behav. Inform. Technol.*, 25(2), 2006.

[23] J. Ling and P. Van Schaik. The effect of text and background colour on visual search of web pages. *Displays*, 23(5), 2002.

[24] S. Lowel and W. Singer. Selection of intrinsic horizontal connections in the visual cortex by correlated neuronal activity. *Science*, 255(5041), 1992.

[25] Y. Lu, B. Tan, and Y. Wang. Web aesthetics: How does it influence the sales performance in online marketplaces. In *Proc. ICIS*, 2013.

[26] R. Maity and S. Bhattacharya. A model to compute webpage aesthetics quality based on wireframe geometry. In *Proc. INTERACT*, 2017.

[27] E. Michailidou, S. Harper, and S. Bechhofer. Visual complexity and aesthetic perception of web pages. In *Proc. SIGDOC*, 2008.

[28] A. Miniukovich and A. De Angeli. Computation of interface aesthetics. In *Proc. CHI*, 2015.

[29] M. Moshagen, J. Musch, and A. S. Göritz. A blessing, not a curse: Experimental evidence for beneficial effects of visual aesthetics on performance. *Ergonomics*, 52(10), 2009.

[30] M. Moshagen and M. T. Thielsch. Facets of visual aesthetics. *Int. J. Hum. Comput. Stud.*, 68(10), 2010.

[31] G. Moss and R. Gunn. Gender differences in website production and preference aesthetics: preliminary implications for ICT in education and beyond. *Behav. Inform. Technol.*, 28(5), 2009.

[32] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. ICML*, 2010.

[33] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE T. Knowl. Data En.*, 22(10), 2010.

[34] I. O. Pappas, K. Sharma, P. Mikalef, and M. N. Giannakos. How quickly can we predict users' ratings on aesthetic evaluations of websites? employing machine learning on eye-tracking data. In *Proc. I3E*, 2020. LNCS 12067.

[35] S.-E. Park, D. Choi, and J. Kim. Critical factors for the aesthetic fidelity of web pages: empirical studies with professional web designers and users. *Interact. Comput.*, 16(2), 2004.

[36] V. M. Patrick. Everyday consumer aesthetics. *Curr. Opin. Psychol.*, 10, 2016.

[37] K. Reinecke and K. Z. Gajos. Quantifying visual preferences around the world. In *Proc. CHI*, 2014.

[38] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proc. CHI*, 2013.

[39] D. Robins and J. Holmes. Aesthetics and credibility in web site design. *Inf. Process. Manag.*, 44(1), 2008.

[40] K. E. Schmidt, Y. Liu, and S. Sridharan. Webpage aesthetics, performance and usability: Design variables and their effects. *Ergonomics*, 52(6), 2009.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. ICLR*, 2015.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015.

[43] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *Proc. ICANN*, 2018.

[44] M. T. Thielsch and G. Hirschfeld. Spatial frequencies in aesthetic website evaluations – explaining how ultra-rapid evaluations are formed. *Ergonomics*, 55(7), 2012.

[45] M. T. Thielsch, J. Scharfen, E. Masoudi, and M. Reuter. Visual aesthetics and performance: A first meta-analysis. In *Proc. MuC*, 2019.

[46] L. Thorlacius. The role of aesthetics in web design. *Nord. Rev.*, 28(1), 2007.

[47] N. Tractinsky. *Visual Aesthetics.* Interaction Design Foundation, Aarhus, Denmark, 2nd edition, 2006.

[48] N. Tractinsky, A. Cokhavi, M. Kirschenbaum, and T. Sharfi. Evaluating the consistency of immediate aesthetic perceptions of web pages. *Int. J. Hum. Comput. Stud.*, 64(11), 2006.

[49] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cogn. Psychol.*, 12(1), 1980.

[50] A. N. Tuch, J. A. Bargas-Avila, and K. Opwis. Symmetry and aesthetics in website design: It's a man's business. *Comput. Hum. Behav.*, 26(6), 2010.

[51] A. N. Tuch, E. E. Presslaber, M. StöCklin, K. Opwis, and J. A. Bargas-Avila. The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. *Int. J. Hum. Comput. Stud.*, 70(11), 2012.

[52] R. W. Veryzer Jr. Aesthetic response and the influence of design principles on product preferences. *Adv. Consum. Res.*, 20, 1993.

[53] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep learning for computer vision: A brief review. *Comput. Intel. Neurosc.*, 2018, 2018.

[54] O. Wu, Y. Chen, B. Li, and W. Hu. Evaluating the visual quality of web pages using a computational aesthetic approach. In *Proc. WSDM*, 2011.

[55] O. Wu, H. Zuo, W. Hu, and B. Li. Multimodal web aesthetics assessment based on structural SVM and multitask fusion learning. *IEEE Trans. Multimedia*, 18(6), 2016.

[56] X. S. Zheng, I. Chakraborty, J. J.-W. Lin, and R. Rauschenberger. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *Proc. CHI*, 2009.