Temporal Visualization of the Trend of Website Design Using Low-level Visual Features (2002-2017)

Ankit Swarnkar, Snehal Vartak, Yingnan Ju, Kunalan Ratharanjan, Yi Bu
School of Informatics, Computing, and Engineering, Indiana University
ankswarn@indiana.edu; snehchem@iu.edu; {yiju,kunarath,buyi}@iu.edu

Abstract—The goal of the study is to investigate the trend of website design over time and to identify the impact of the change in website design on technology and culture. We use computer vision techniques to extract features from 24k+ figures within five dimensions, color, symmetry, complexity, texture, and luminance. The dynamic visualization is based on d3.js. Several static screenshots are shown in Figure 2 as examples. Some validations are implemented for the visualizations.

Index Terms—Website evolution, design, visualization, computer vision, d3, JavaScript

1 Introduction

With over 25 years of history, the web itself has become a significant cultural artifact [1]. In these years, websites give owners direct control over the message that they aim to deliver to target customers in a cheap, flexible, and "24-7-365" way and are thus regarded as powerful tools for businesses. A well-designed website is found to attract more potential consumers [2]. Therefore, examining how the design of websites evolves is of importance to the industries.

In this project, we study how website design has changed over time and how these reflect changes in terms of culture, technology, and aesthetics. This project is designed to blend the computer vision and information visualization so that we can understand the importance of low features of an image (website screenshot) and how the importance of distinctive features evolved over time. We use features including color, symmetry, complexity, texture, and luminance to show how the design of websites evolve these years. Further, this project includes the visualization part which enables us to identify the best visualization type and technique to represent the insights to the user in an effective way. Feature extraction allows us to use the knowledge that we gained in CS B657 class about computer vision and expand our computer vision knowledge for achieving the goal of this project.

2 Related Work

By achieving the goal of the project, we will be able to see the changes in the website design over 15 years. This analysis would lead to the prediction of future changes in the website design, which could serve as a reliable source for several purposes. From the research perspective, previous studies concerning website evolution mainly focused on two perspectives, *empirical*- and *theory-oriented*.

The empirical-oriented branch of studies provided heuristics or checklists for the design of websites without many theoretical foundations [3, 4]. For example, [5] revealed how color affects trust or satisfaction from the users' (viewers') part. [6] identified several underlying dimensions of effective website design and "provided insight into design characteristics" (p.787). [7] mined the visual evolution in 21 years of web design by using deep neural networks, but their features used are limited. Similarly, [8] tried to understand the aesthetic evolution of websites by providing a feature-based empirical study.

The second branch showed more theoretical views. For instance, [9] presented a two-factor model for website design and evaluation and found that hygiene and motivator are two important aspects. Specifically, hygiene factors make a website functional and serviceable; a website lacking hygiene factors will dissatisfy users. Nevertheless, motivator factors facilitate users to have more satisfactions. [10] concentrated on the tasks in information retrieval and showcased usability studies on several websites of some big companies. [1] summarized the history of website design by providing a quantitative empirical study.

3 Methodology

3.1 Dataset

Screenshots of hundreds of websites over a period of almost 15 years are used as input dataset for the current analysis. Web scraper was used to obtain this dataset from the Internet Archive's Wayback Machine. The dataset consists of snapshots of 100 different websites, such as apple.com and google.com, at different time from 2002 to 2017. Each website has more than 200 snapshots images available and we thus have a total number of 24174 images as our raw dataset.

Each image is stored as a matrix containing a considerable number of pixels; every pixel is represented as a four-dimension vector showing the corresponding values on the colors of red, green, and blue, as well as a parameter indicating transparency (alpha), ranging from 0 to 255. For instance, a purely white pixel with absolute opacity is represented as [255, 255, 255, 255]. More descriptive metadata of the dataset can be found in the online supplement material.

3.2 Methods

Figure 1 shows the flow chart of our method, in which we can see that we first filter the raw dataset both automatically (see in Section 3.2.1) and manually (Section 3.2.2), and then extract features based upon five dimensions from images in the filtered dataset. The visualization using d3.js (Section 3.2.3) is then implemented. Finally, we communicate with the sponsor of this project, validate our visualization results, and do some predictions based on our built model (Section 3.2.4).

3.2.1 Data filtering

Automatic and manual data filtering are both implemented in this procedure.

Automatic data filtering: The automatic data filtering is mainly based on two indicators, the ratio of transparent pixels and the size of the image files. First, we find all images that contain transparent pixels (alpha = 0), calculate the ratio of transparent pixels and all pixels, and remove the images with the ratio greater than a ratio threshold (50%) and with file size smaller than a size threshold (25 KB). The reason why some screenshots of websites are stored in as a small file and/or they feature a high ratio of transparent pixels is because the snapshotting on a certain website had been done before the whole website was fully rendered (for example, the 2008 Website of "factset.com"; see more details in the online supplemental material). For those whose file size is extremely small (< 16 KB), we also remove them from our dataset.

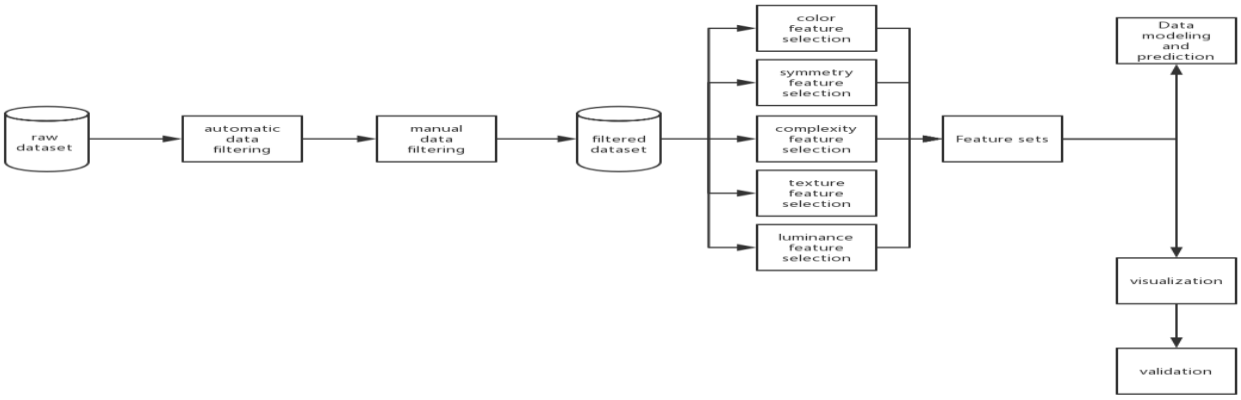

Figure 1. Flow chart of our method.

Manual data filtering: We then manually examine our automatic filtering results and correct them if possible. A typical case is Google's website. By using the aforementioned automatic data filtering approach, all images of Google's websites are deleted from our raw dataset because most parts of the images are white background and thus these images have very small image size (less than 1/8 of the average file size).

3.2.2 Feature extraction

Computer vision techniques are employed here to extract features from our filtered dataset. Five dimensions of the features are selected, namely color, symmetry, complexity, texture, and luminance [8-10]. Each website in each month has values of the above-mentioned features.

Color features extraction: The color features are extracted from five main portions of the website: center, top right, top left, bottom left, and bottom right, as shown in Figure 2. Images were transformed from original RGB space to perceptual HSV space. For each portion, we extract 4096 (=16*16*16) features by quantizing the H, S and V color space into 16 bins which resulted in 20480 features. To visualize the color features, we perform a principal component analysis (PCA) on each regions' 4,096 features of a region which resulted in one feature for each of the five portions of the website.

Symmetry features extraction: We mainly consider three types of symmetry features, left-right (horizontal symmetricity) [11], top-bottom (vertical symmetricity) [11], and central (diagonal) symmetries. All of these symmetries are determined by the difference between the raw matrix indicating a given image mathematically and the transformed matrix indicating the symmetric version of the corresponding image. Note that these differences are normalized into [0,1] for better visualization purposes. A greater value in these features shows higher symmetric characteristics in a given image.

Figure 2. An example of our extracted color features.

Complexity features extraction: Three indicators are considered in the set of complexity features, including Quadtree value, compression ratio, and information entropy. Quadtree [12] is an algorithm that majorly divides a given image into four different parts, each of which comprises a quarter of the image. As an iterative algorithm, Quadtree stops when meeting one of two thresholds set in advance, the number of pixels included in a quarter and the standard deviation of the color in a quarter. Then count the number of Quadtree leaves, which is an important factor to measure the complexity of an image. Meanwhile, the compression rate of an image is also used as an indirect indicator for measuring the complexity of an image. An image with higher compression ratio is considered as less complexity, and *vice versa* [13]. Information entropy [14] is another indicator employed here; an image with higher information entropy is regarded as higher complexity. All of these three features are normalized into [0,1].

Texture features extraction: Texture can be quantified in terms of texton which refer to the smallest patch of an image with a similar pattern. Multi Texton histogram (MTH) integrates the advantage of co-occurrence matrix using histogram and can be considered as a generalized visual descriptor. MTH were prepared for each image using 82 bins in RGB space. We applied PCA on the extracted 82 values to get a single feature for visualization. The feature was scaled to [0,1] using robust median scaling.

Luminance features extraction: Luminance is a photometric measure of intensity per unit area of light travelling in a given direction. We explored three perspectives of Luminance: relative luminance [15], color brightness and perceived brightness [16]. Relative luminance are normalized values from 1 to 100 with white light as a reference. [15] proposed that relative luminance can be created from RGB space by first converting gamma compressed RGB value to linear RGB value and applying following formula:

relative luminance = 0.2126*R + 0.7152*G + 0.0722*B

Color brightness can be calculated by converting RGB value into YIQ values using the below formula:

YIQ = (R*299 + G*587 + B*114) / 1000

[17] suggested an alternative view of perceived luminance value which can be calculated using the following formula:

L = sqrt (0.299*R^2+0.587*G^2+0.114*B^2)

All the three features were extracted and normalized into [0,1].

3.2.3 Visualization

The goal of the creation of the visualization is to identify the patterns of website design has changed over time with the given low-level features. To prepare the data for visualization, the data are aggregated over the year column, and the mean values for each portion of the website for each year is obtained. We here employ d3.js [18] to show the visualization of website design evolution in a dynamic way. Specifically, the names of websites and the features in a certain dimension (e.g., top in the dimension of color) are set as two dropdowns beside the dynamic trend lines.

3.2.4 Modeling and prediction

We also build a simple model to see how these features correlate with each other and try to do some predictions for them.

4 Results and Discussion

In this section, we showcase several visualization and model results based on each feature to indicate the website design evolution between 2002 and 2017 and describe how they changed over years. We tried using two model approach: *Model 1 – Year Prediction and Model-II-Era prediction* using similar approach like Doosti[1] and got better results.

Color features: Figure 3 shows how color features changed from 2002 to 2007 taking adobe.com as an example, in which we can find that the color feature of the web page is relatively low. Most of the web pages did not contain the colorful stuff (mostly contents and paragraphs). After 2008, there is increasing changes, as we can see in the color feature up to 2014. Then in the 2015, there is a sudden decrease in the color feature, yet it slightly increased. In 2014, a prominent number of web pages contain color features. As a section wise most of the times "center part of the web page" contain more colorful features than other sections; "top right part of the webpage" section mostly contain the contents than colorful features in most of the years. Our fitted model suggested that color can be a prominent feature to distinguish between years.

Symmetry features: From 2002 to 2017, different websites have a different trend of symmetry features. For example, cigna.com (see in Figure 4) have very symmetry webpages from 2008 to 2011 but after 2011 the symmetricity dropped to the level before 2008. Adobe.com has a trend of reducing the symmetricity but google.com always has a constant high value of symmetricity. Overall, the symmetry features of all websites are getting lower year by year, especially the vertical symmetry feature. The lower value of symmetry features could also reflect that most websites have a more complex layout of the page with more multimedia contents. Our fitted model suggested symmetry is not an important feature for year but can help for identifying the era.

Complexity features: Figure 5 shows the mean value of all websites' changes over years in terms of complexity features. From 2002 to 2017, most websites have the trend of a more and more complex webpage. Therefore, the overall trend is the rise of the complexity feature year by year, especially after 2006. Another proof of the complexity of the web pages is the animated gif shown in the visualization. The mean image of all websites in one year is getting darker from 2003 to 2017 and that means more and more areas of the page is utilized to show more contents. Our fitted model suggested complexity can be an important feature for identifying the website before 2011 however after that the feature is not useful.

Texture features: There is a clearly increasing trend from 2002 to 2017 (See figure 6) though some jiggers among texture features. This shows that designer paid more attention on increasing the texture space. High value of texture may be explained with more inclusions of graphics and pictures that can be the reason of this clear trend of increase. Our model also suggested that texture is an important feature for the evolution

Luminance features: There is a subtle decrease in the luminance calculated using relative luminance and color brightness (Figure 7) which can be due to the reduction of white spaces and inclusion of eye suitable colors as they tend to have less luminance. The two features work well in non-perceptual color space, however, fails to demonstrate the human perception. Regarding the human perception of luminance, there is a slow increase in term of perceptive luminance that is closer to human way of perception. Our model was not able to capture the trend.

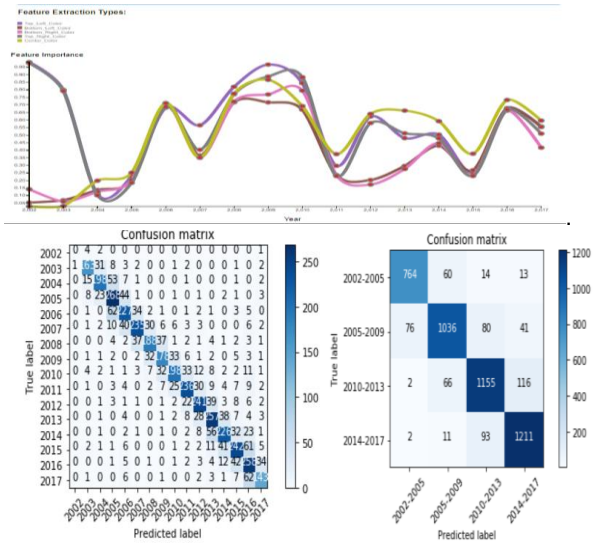

Figure 3. Visualization and model for color feature

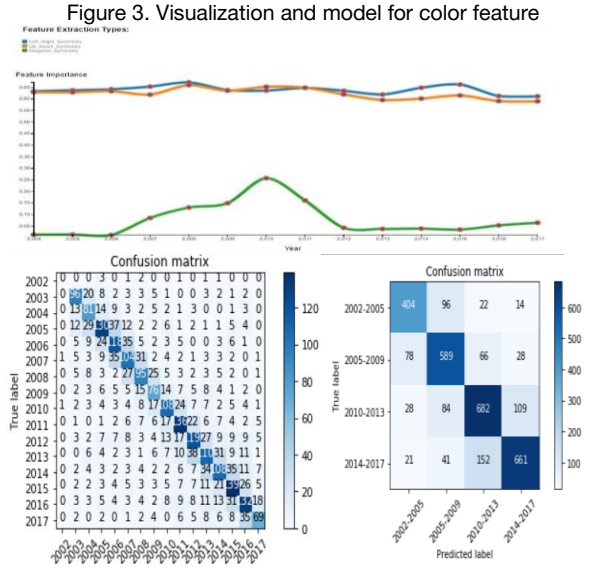

Figure 4. Visualization for symmetry features: An example of cigna.com.

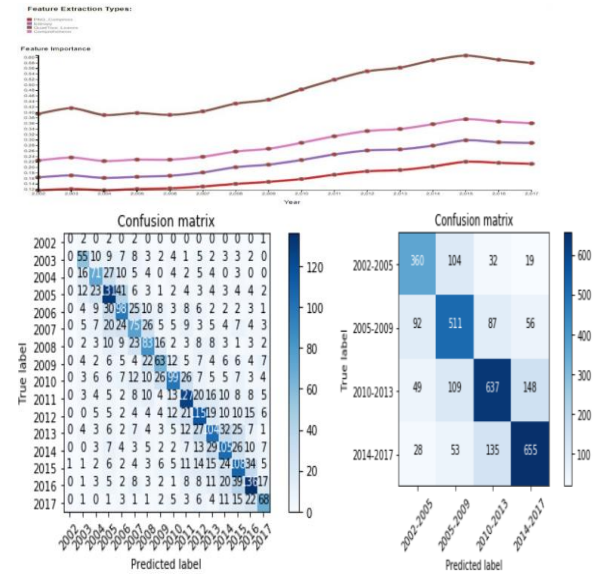

Figure 5. Visualization and model for complexity features (all websites).

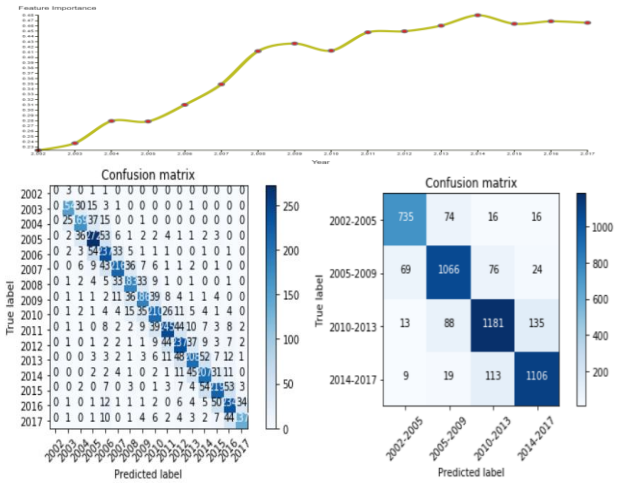

Figure 6. Visualization and model for texture features

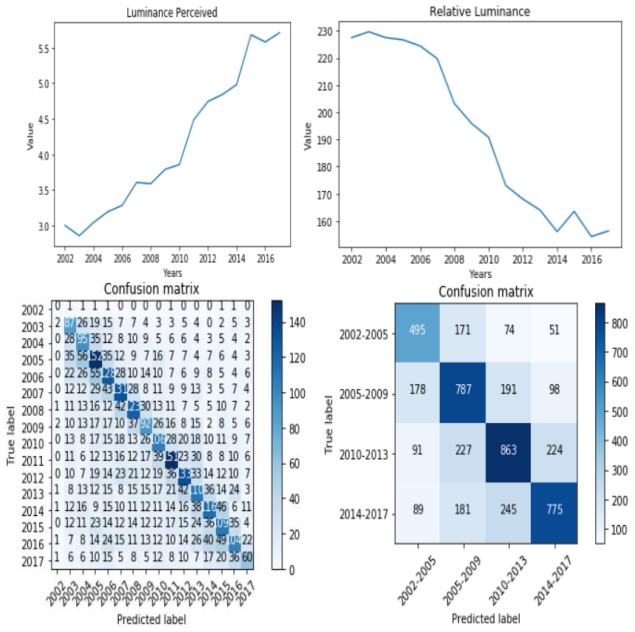

Figure 7. Trend and model for Luminance features

The model results are shown in the table below. We tried using KNN, Logistic, RF and NN(2-layers) and RF gave us better result for both model. The model was tuned separately for each set. Each set of features was divided into test and train set with 80-20 split and accuracy score below is reported on the test set.

Feature	Year Accuracy (%)	Era Accuracy(%)
Color	68.7	89.8
Texture	67.1	86.2
Complexity	51.7	72.8
Luminance	35.8	61.6
Symmetry	52.7	75.5

5 Conclusions

The goal of the study is to investigate the trend of website design over time and to identify the impact of the change in website design on technology and culture. We use computer vision techniques to extract features from 24k+ figures within five dimensions: color, symmetry, complexity, texture, and luminance. The dynamic visualization is based on d3.js and shows superior performance. Some validations based on communications with the clients are also implemented for the visualizations.

6 Supplemental Material

Due to the page limits, we provide the supplemental material online:
http://cgi.soic.indiana.edu/~ankswarn/deepweb/index.html.

Acknowledgments

We wish to thank David J. Crandall for his guidance and suggestions on improving this project. We are also grateful to Katy Börner, Michael Ginda. We also like to thank TAs who helped us with their great feedbacks.

References

[1] Doosti, B., Crandall, D.J., & Su, N.M. (2017, June). A Deep Study into the History of Web Design. In Proceedings of the 2017 ACM on Web Science Conference (pp. 329-338). ACM.

[2] Liang, T.P., & Lai, H.J. (2002). Effect of store design on consumer purchases: an empirical study of on-line bookstores. Information and Management, 39(6), 431-444.

[3] Flanders, V., Willis, M. (1998). Web Pages that suck. San Francisco, CA: SYBEX Inc.

[4] Simon, S.J. (2000). The impact of culture and gender on web sites: an empirical study. ACM SIGMIS Database: The Database for Advances in Information Systems, 32(1), 18-37.

[5] Cyr, D., Head, M., & Larios, H. (2010). Color appeal in website design within and across cultures: A multi-method evaluation. International journal of human-computer studies, 68(1-2), 1-21.

[6] Rosen, D. E., & Purinton, E. (2004). Website design: Viewing the web as a cognitive landscape. Journal of Business Research, 57(7), 787-794.

[7] Jahanian, A., Isola, P., & Wei, D. (2017, May). Mining Visual Evolution in 21 Years of Web Design. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2676-2682). ACM.

[8] Chen, W., Crandall, D.J., & Su, N.M. (2017, May). Understanding the Aesthetic Evolution of Websites: Towards a Notion of Design Periods. In CHI (pp. 5976-5987).

[9] Zhang, P., & Von Dran, G.M. (2000). Satisfiers and dissatisfiers: A two-factor model for website design and evaluation. Journal of the American Society for Information Science, 51(14), 1253-1268.

[10] Spool, J., Scanlon, T., Schroeder, W., Snyder, C., DeAngelo, T. Web site usability—A designer's guide. San Francisco, CA: Morgan Kaufmann Publishers, Inc.

[11] Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013, April). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 2049-2058). ACM.

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

[13] Jones, G.A., Paragios, N., & Regazzoni, C.S. (Eds.). (2012). Video-based surveillance systems: computer vision and distributed processing. Springer Science & Business Media.

[14] Yang, J., & Yang, K. (2004). Adaptive detection for infrared small target under sea-sky complex background. Electronics Letters, 40(17), 1083-1085.

[15] Poynton, C. (2012). Digital video and HD: Algorithms and Interfaces. Elsevier.

[16] www page of Web Accessibility Initiative working draft, https://www.w3.org/TR/AERT/#color-contrast accessed April 21, 2018.

[17] Finley, Darel Rex. "HSP Color Model — Alternative to HSV (HSB) and HSL." The WWW Homepage of Darel Rex Finley. 2006. http://alienryderflex.com/hsp.html (accessed November 25, 2008).

[18] https://d3js.org