Visualizing the Evolution of Website Design

Surbhi Paithankar
Data science(Grad)
spaithan@iu.edu

Apurva Gupta
Data science(Grad)
guptaapu@iu.edu

Shyam Narasimhan
Data science(Grad)
shynaras@iu.edu

We are working with prof David Crandall. Our work is an extension of his previous work.

OBJECTIVE AND MOTIVATION

Visual features of websites are important signals to study different areas of humanity like Technology, aesthetics, cultures and industries. Paintings can be used to study an era's social norms and culture. Similarly, analyzing the visual appearances of websites can help us know the global changes of websites.

The goal is to find out how much information about cultural patterns are encoded solely in the visual appearances of sites. Thus, visual designs of websites could provide us an understanding of evolution of web i.e. changes in visual aesthetics, role of technology, cultural preferences and technical innovation.
This motivates us to study these visual changes happening over time.

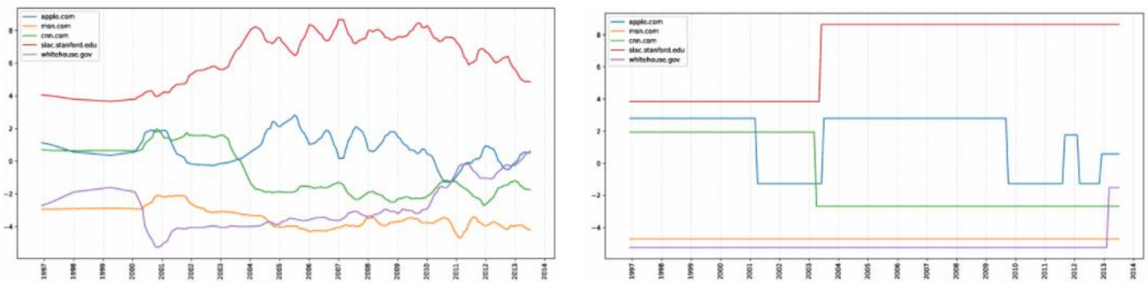
BACKGROUND

PREVIOUS WORK

In previous work by Reinecke et al., they have used low level features for quantifying the visual properties of websites such as color distribution, amount of white space and structure of page layout and conducted a survey to see that how users rate these features on websites of different years. This was one of the first steps in understanding the cultural patterns via analysis of visual designs of websites.

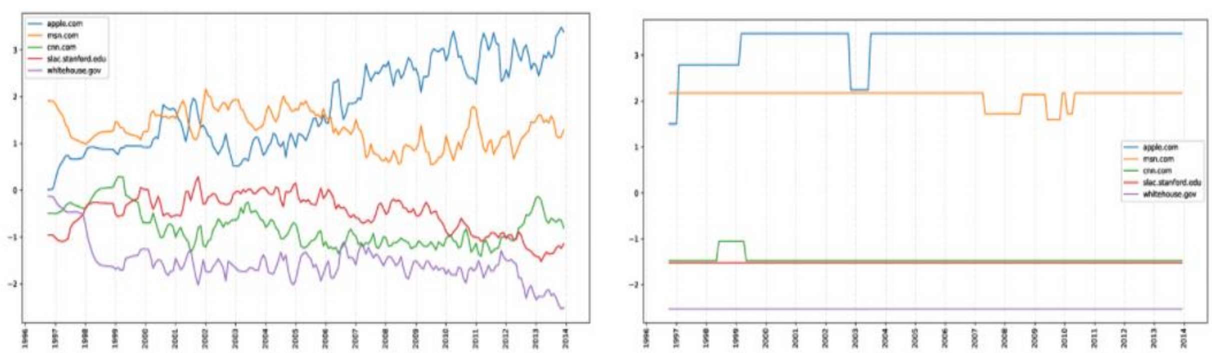
Previous work done by Prof David Crandall and Bardia Doosti was focused on using convolutional neural networks to recognize the eras and genres of web pages. The results obtained showed that classification of websites by genre gave an accuracy of about 4 times the baseline and recognizing website era by 2.5 times the random baseline. These results gave us a mechanism to measure the similarity of a given page to a genre and similarity between a given page to its era. These results were really fascinating which showed that convolutional neural networks can help us in understanding the visual patterns on web pages. Convolutional neural networks were further used to measure the visual design similarity between two pages i.e. how similar is website 1 to website 2.

For example, in the below visualization, the similarity of Indiana.edu is compared against each canonical site.



From the visualizations above, it is apparent that Indiana.edu is very closely related to slac.stanford.edu, which is another educational website.

Finally, all the sites were merged to plot the confidence of the network to see how web design has changed over time.



As we see in this graph, websites were detected to be very similar to msn.com until 2006, when they were detected to be more and more like apple.com. As we can also see in the HMM generated graph, the other three websites didn't seem to change a lot in these years.

Also, Kumar et al. introduced a new platform which is called Webzeigeist to generate features like ratio, dominant colors and number of words. These features were captured from the HTML Code and thus it can capture these features only roughly. Moreover, many papers have extracted the text i.e. the HTML code from the web pages and used that for classification of genre and other properties of web pages.

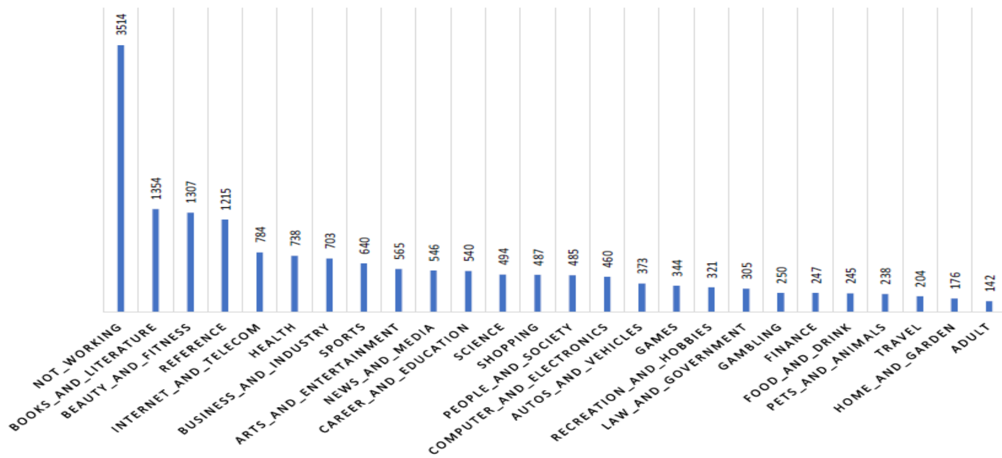
Our Approach

We have combined Kumar et al. and Prof David Crandall's work, to extract features like dominant colors, colorfulness and number of words from the web page snapshots instead of HTML text. This helps us in capturing these features accurately.
We have showed that how these features are changing with respect to time and Genre. Moreover, we have improved the accuracy of classification by genre from 4 times the random baseline to about 5.5 times.
We have created Tableau Dashboard to visualize the changes in visual design.

DATA DESCRIPTION

We have two kinds of datasets available: -

1. **Snapshots of websites organized according to their genre**
CrowdFlower's URL categorization dataset has been used. It consists of more than 31,000 URL domains along with their manually annotated genre. Data is classified into 26 genres (News, sports, business etc.).
These URLs are used to fetch their web snapshots using PhantomJS, Webkit API. The snapshots have a resolution of 1200 * 1200.



It shows the frequency of webpages in each genre.

2. **Historical Snapshots of websites over time**
The snapshots were collected from the Internet Archive. However, this data set is sparse since many well-known websites did not exist before 1990s. We have 7303 screenshots from 35 chosen websites. This has been collected from archive.org that span from 1996 through 2013.
Using PhantomJS, the web pages were converted into images of resolution 1200 * 1200.

The data was provided to us by Prof David Crandall.
We manually annotated this data according to 4 Genres: Media, Internet and Telecom, Education and Technology.

Genre	**No. Of websites in that Genre**
Media	5
Education	13
Internet and Telecom	8
Technology	9

We have divided years into 4 eras:
- 1996-2000
- 2001-2004
- 2005-2008
- 2009-2013

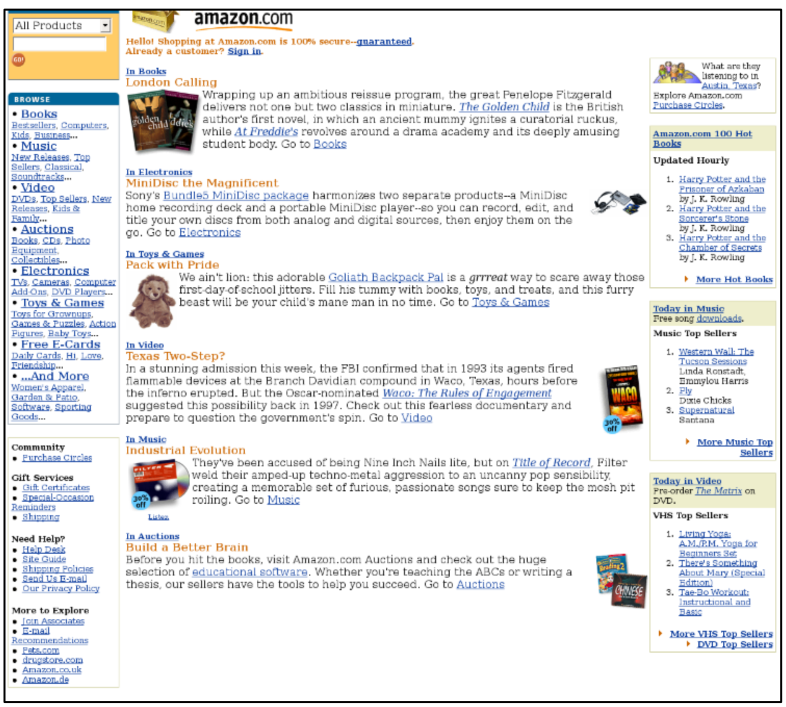
METHODOLOGY

We performed the following tasks:

1. Counted the average number of words on webpages according to website and year.
2. Found the top 6 dominant colors on a webpage using clustering.
3. Found colorfulness measure according to website and year
4. Classified images according to era using convolutional neural networks
5. Classified images according to genre using convolutional neural networks

TASK 1: Count the average number of words on webpages according to website and year

- Optical character recognition is the process of recognizing text from images. Various models have been developed over the years, ranging from probabilistic models to graphical models like Hidden Markov Models, and finally to Convolution Neural Networks for recognition of texts.
- We wanted to see the trend of count of words across eras for different websites.
- We used in built functions of Py-Teserract library to recognize the texts from any given image.
- The words were recognized by using the separators as space, tab and line.

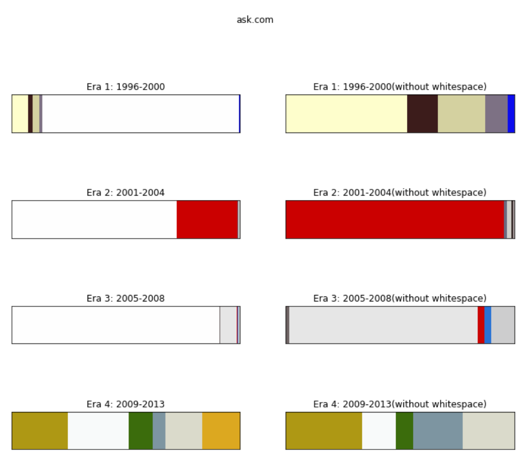


Average number of words on this website snapshot are 601.

TASK 2: Finding the top 6 dominant colors on a webpage using clustering

- We used OpenCV for extracting features and K -means clustering algorithm to cluster a web page into 6 clusters.
- The goal of k means clustering algorithm is to partition n data points into k clusters. Each of the data points will be assigned to a cluster with the nearest mean i.e. centroid. Data points inside a cluster tend to be more similar to each other than data points that belong to different clusters.
- Here, we have RBG image. We will cluster the pixel intensities of RGB image.
- Our data points are M×N pixels and we will cluster them using k-means. Points that belong to a given cluster will be more similar in color than pixels that belong to different cluster.
- We used sklearn's k mean package to cluster. The centroid gives the dominant color values.
- Webpages have a lot of whitespace content. Therefore, we found the dominant colors excluding white color.
- We plotted bar graphs for percentage of pixel values in each centroids/dominant colors.
- We applied k means on images of each era.

Here is a screenshot of dominant colors of ask.com for each era.



INSIGHTS
We can observe that more sophisticated colors are being used in later eras. Also, we observed that use of grey shades has increased over time. In left side of diagram, we can see that pixels of white color have decreased over time.
Also, the distribution of colors looks to be great in later eras.

TASK 3: Find colorfulness measure according to website and year

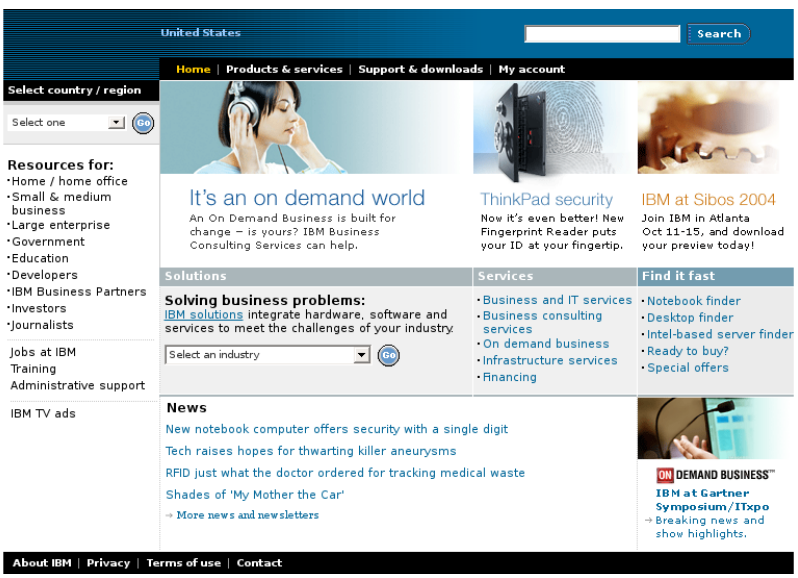
- Calculating colorfulness metric of a webpage is useful to evaluate the 'aesthetic qualities' of a webpage.
- We implemented Hasler and Susstrunk's paper from 2003,'Measuring colorfulness in natural images'.
- We used OpenCV and Python for implementation.
- Image colorfulness metric is defined as follows:

$$Rg = R - G$$
$$Yb = \frac{1}{2}(R + G) - B$$
$$\text{Metric} = \sigma_{rgyb} + 0.3 * \mu_{rgyb}^2$$

$$\text{Where } \sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2}\ ;\ \mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2}$$

- Based on the metric scores, we can classify into one of the following classes:

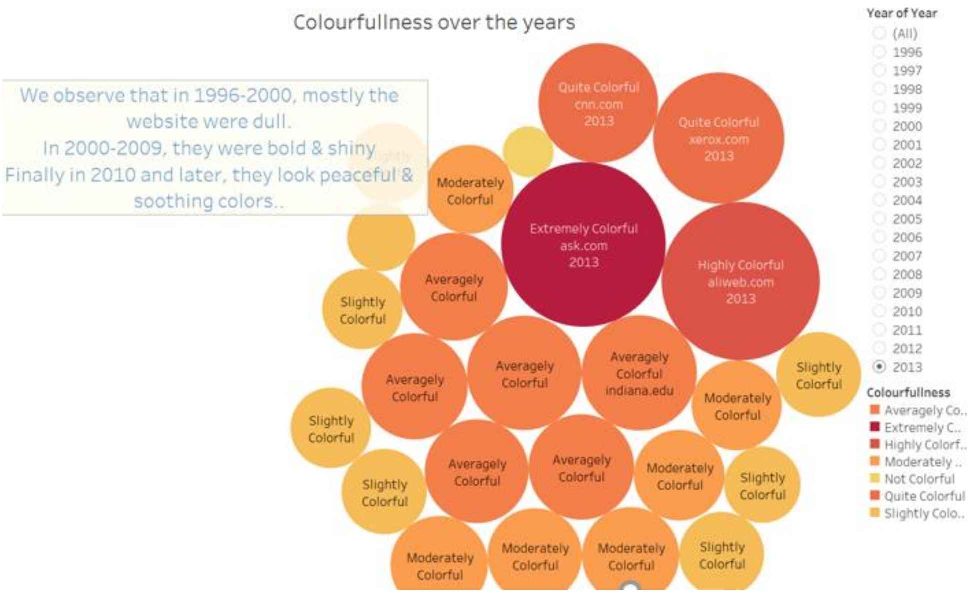
Attribute	Score
Not colorful	0
Slightly colorful	15
Moderately colorful	33
Averagely colorful	45
Quite colorful	59
Highly colorful	82
Extremely colorful	109



This is IBM website from 2004.
Colorfulness score: 25.15
This suggests that the webpage is moderately colorful.

INSIGHTS

We obtained colorfulness metric for websites and different years. It gave us insights that the web designs in early years used to be less colorful/slightly colorful. In mid years, the webpages used to be highly/extremely colorful. The trend is changing again, and all websites are becoming averagely/moderately colorful.

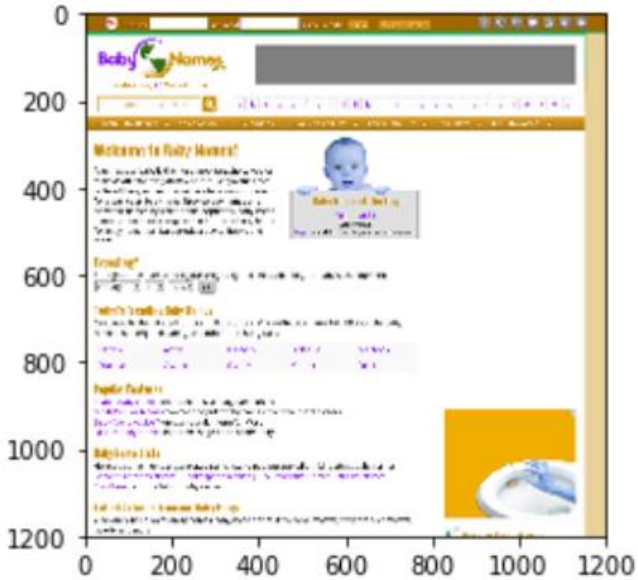


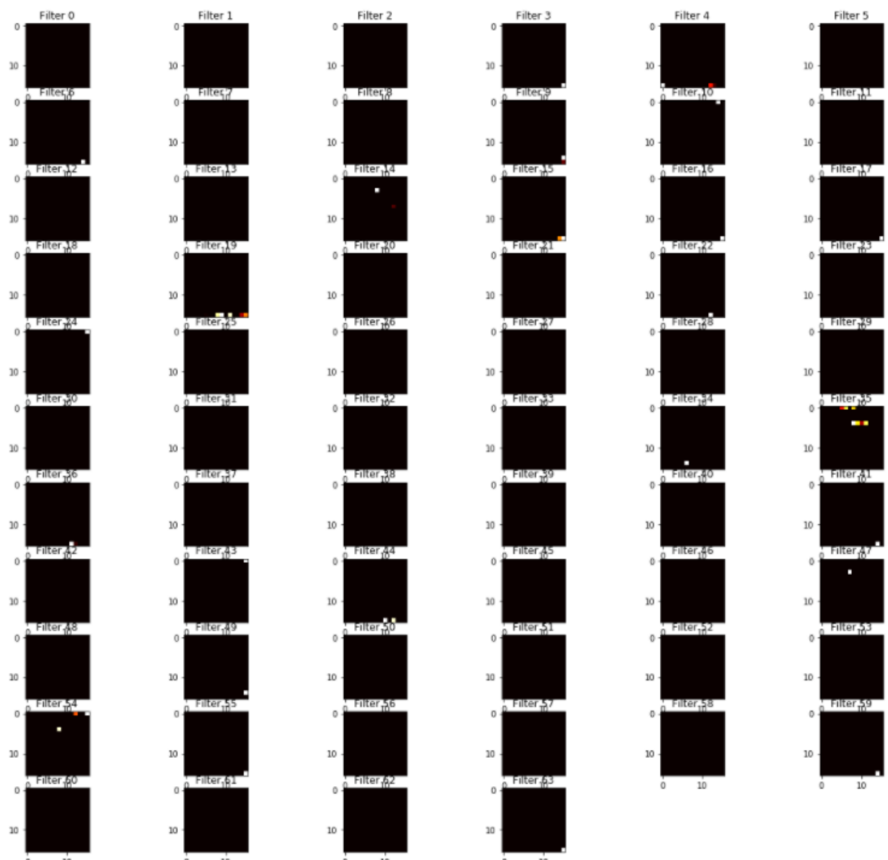
In 2013, all websites are becoming moderately/averagely colorful.

TASK 4: Classification according to genre using convolutional neural networks

- We implemented 2-layered Convolutional Neural Network using tensor flow was used to classify the images into their corresponding genres.
- Data Preparation:
  1. Performed stratified sampling and cross-validated the images into 50-50% train & test split.
  2. Resized image from (1200,1200) to (32,32) for reducing the computational time and complexity.
  3. The images are available that belong to either of the 26 genres.
- **Network Architecture:**
  We used 2 layered convolutional layer followed by full dense layer in our network.
  1. Activation Functions used:
     a. ReLu function
     b. Softmax Layer
  2. Optimizer used: Adam Optimizer
  3. Holding Probability: 0.5

- We displayed a random image and visualized using heatmaps. These heatmaps help us focus on the areas where the various filter paid attention for performing the classification.



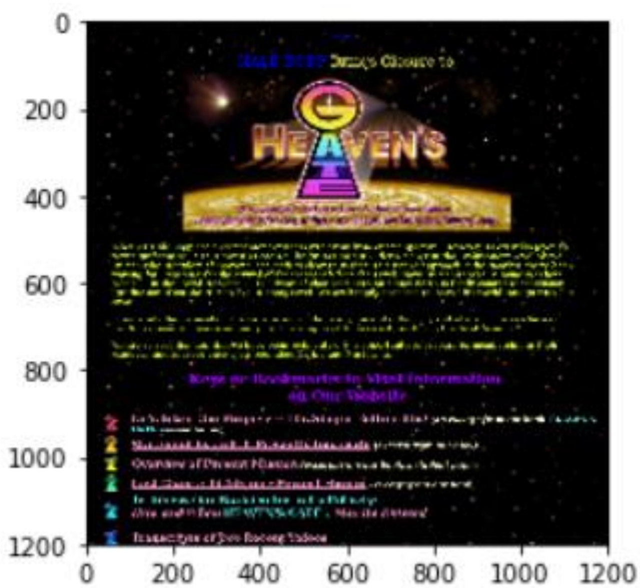


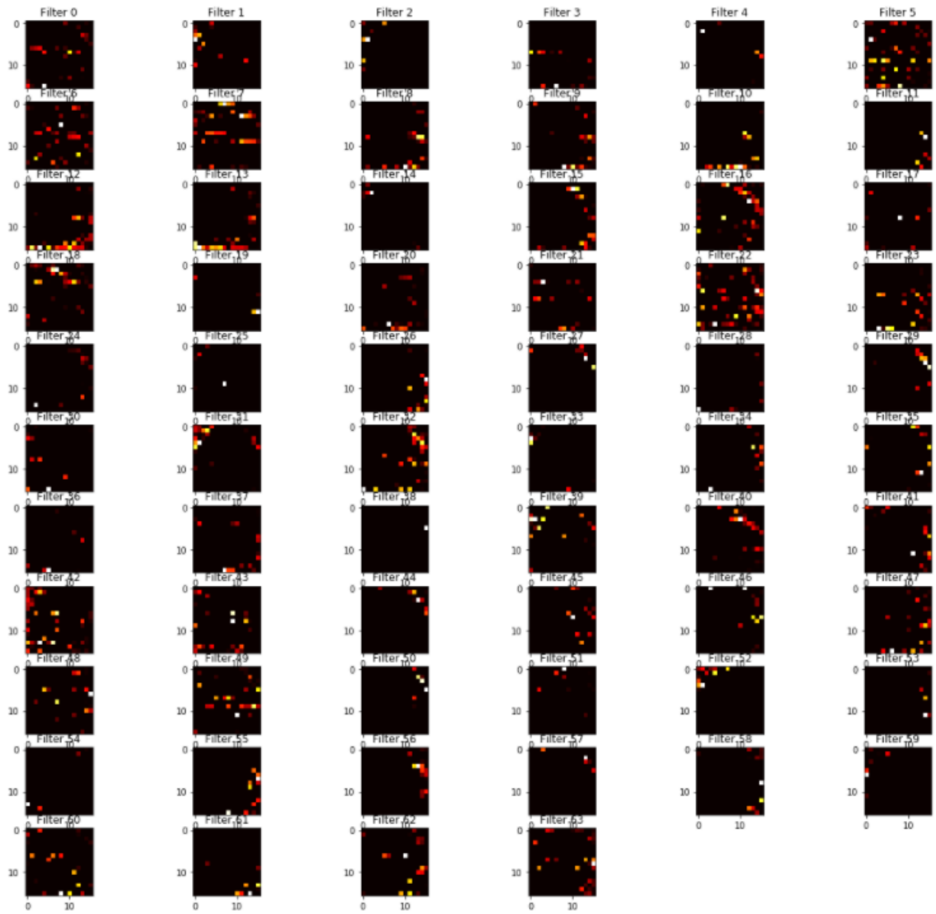
This shows the different parts of image, that convolutional neural network considers important for classification.

TASK 5: Classification according to era using convolutional neural networks

- We implemented 2-layered Convolutional Neural Network using tensor flow was used to classify the images into either of the four different eras.
- Data Preparation:
  1. Performed stratified sampling and cross-validated the images into 80-20% train & test split.
  2. Resized image from (1200,1200) to (32,32) for reducing the computational time and complexity.
- **Network Architecture:**
  We used 2 layered convolutional layer followed by full dense layer in our network.
    4. Activation Functions used:
        c. ReLu function
        d. Softmax Layer
    5. Optimizer used: Adam Optimizer
    6. Holding Probability: 0.5

- We displayed a random image and visualized using heatmaps. These heatmaps help us focus on the areas where the various filter paid attention for performing the classification.





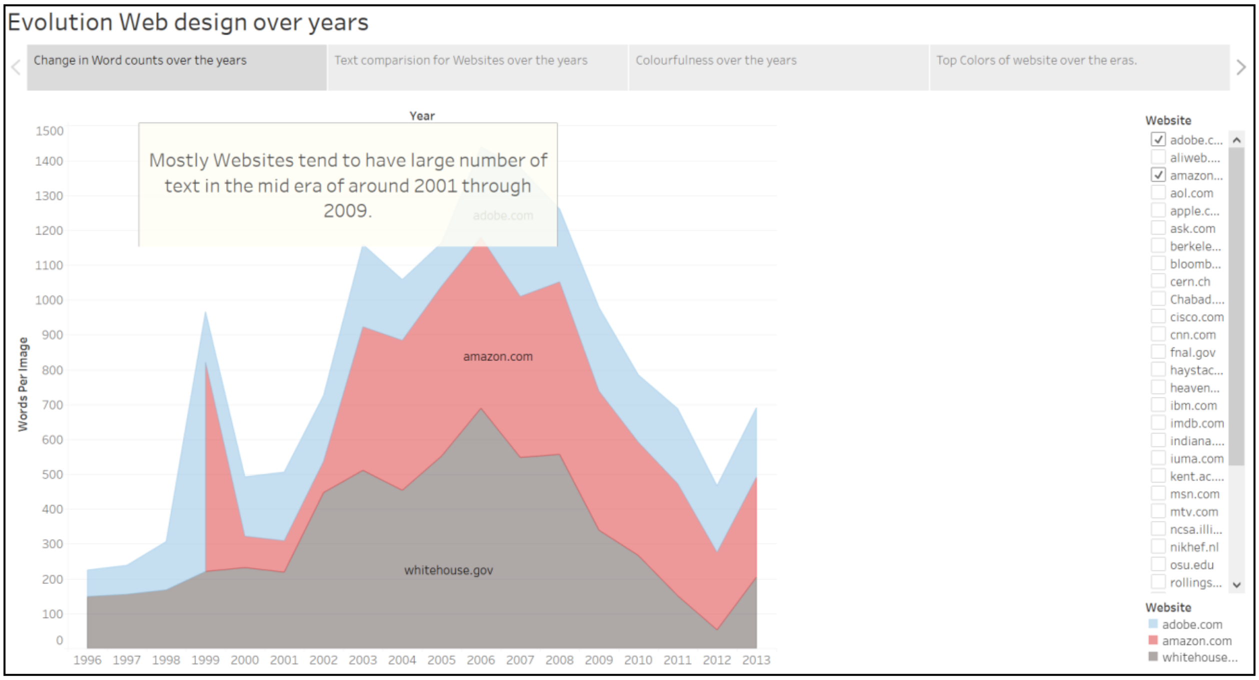
This shows the different parts of image, that convolutional neural network considers important for classification.

OUR VISUALIZATIONS AND RESULTS

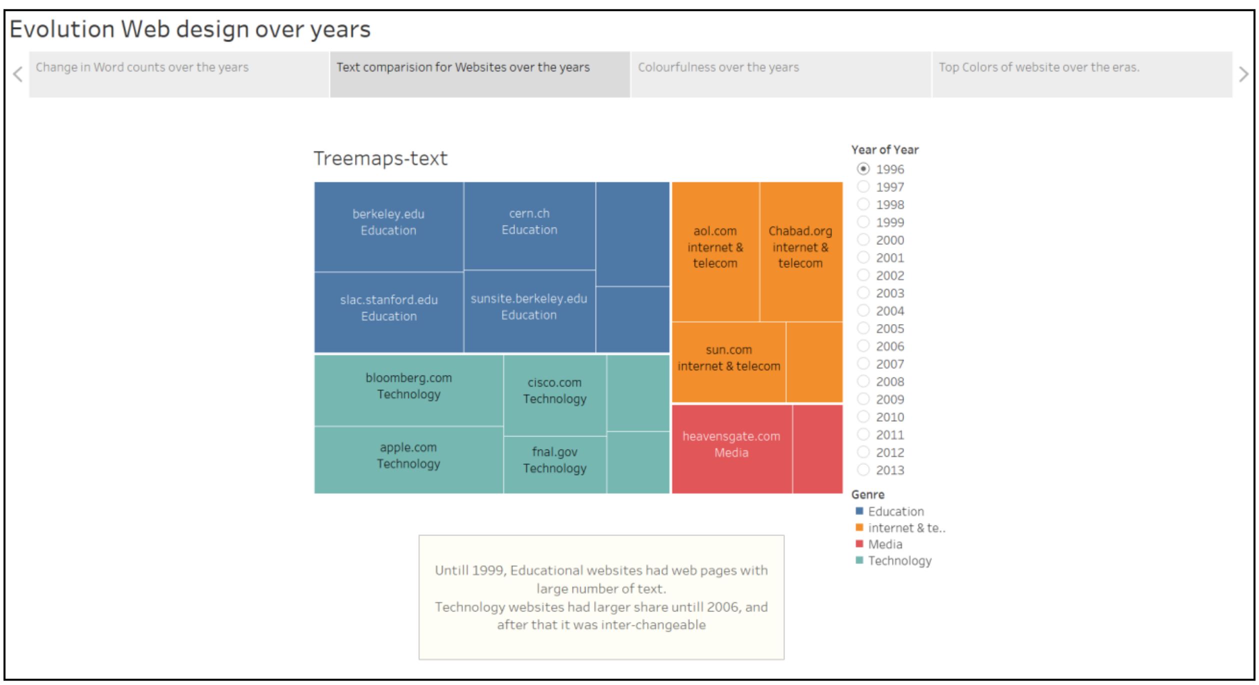
We created a dashboard story with 4 dashboards. The snapshots and explanation of visualizations is as below:

The dashboard can be see using the below mentioned link:
https://public.tableau.com/profile/surbhi.paithankar#!/vizhome/UnderstandingtheevolutionofWebdesignsover18years/OurStory

NUMBER OF WORDS AS FEATURES

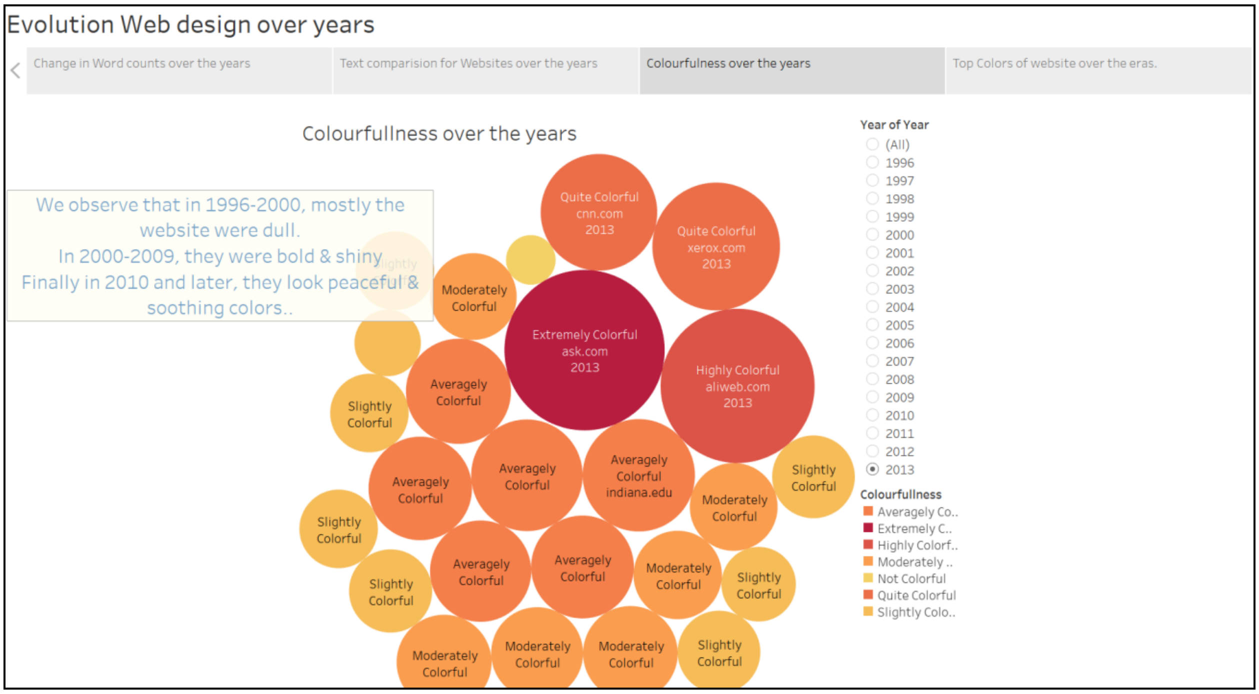


This visualization shows the change in number of words over year. We can see that the number of words on websites used to be very less in initial eras. But in mid years 2001 to 2009, the number of words on website increased drastically. The trend is again changing, and the number of words on web page are decreasing.



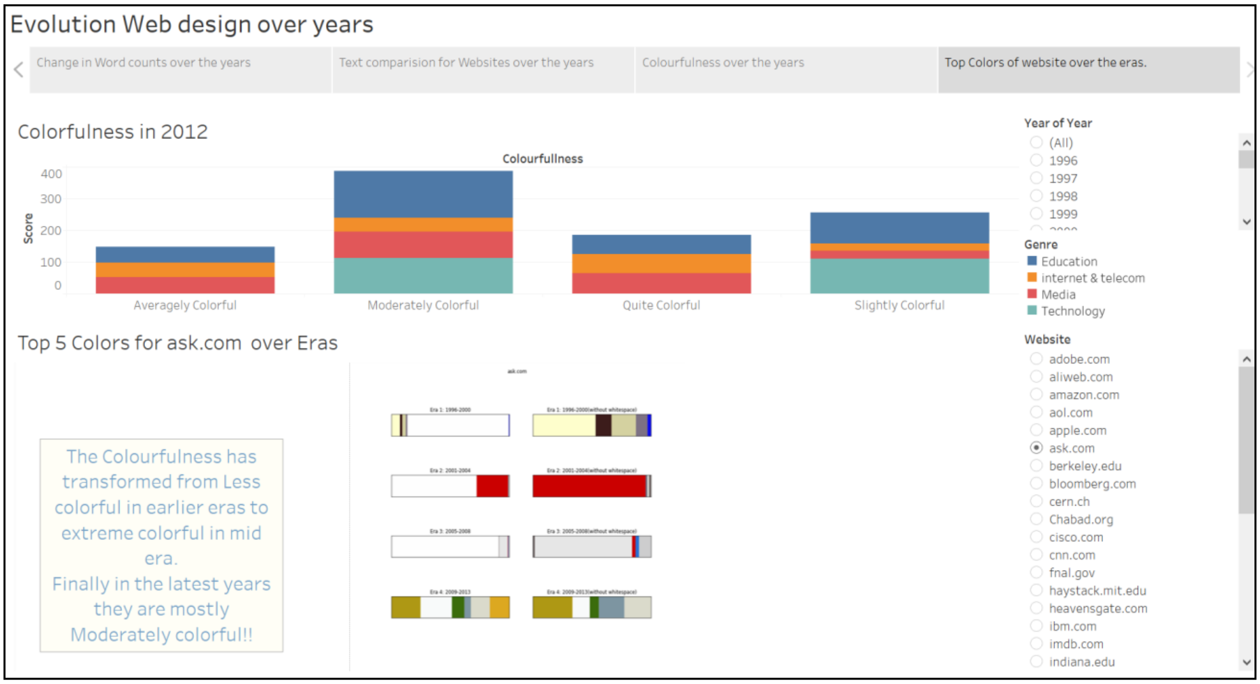
This visualization shows the treemap according to 4 genres. It shows that which genre used to have high number of words compared to others. In early years, Education websites used to have more number of words but in later years, technology websites had more number of words.

COLORFULNESS MEASURE



We have calculated colorfulness using our metric. We have showed the metric value for different websites. This helped us gain insight that in 1996, most websites used to be in category: not colorful or slightly colorful. But as we reached in mid eras, the websites used to be highly colorful. Now, again the websites are becoming averagely/moderately colorful. This shows that colorfulness has increased and then again decreased.

DOMINANT COLORS IN AN IMAGE



This visualization shows bar graphs to visualize colorfulness over Genre. It shows that in 2012, Education websites used to be more colorful than other genres.
The below visualization shows the top 6 colors on each website for all the four eras. We can look at the trends. Earlier eras, used dark colors like red. The use of soothing colors has increased over the years. We also observed that the use of grey shades has increased a lot. Also, the white color is reducing, and it has been replaced by grey shades.

RESULTS FOR CLASSFICATION BASED ON GENRE

We measured our performance using the accuracy of our classifier.

Accuracy = No. of correctly classified image/ Total number of images * 100

$$\text{Accuracy} = \frac{\text{No. of correctly classified image based on the genres}}{\text{Total number of images}} * 100$$

Accuracy = ~21%.

We were able to improve the accuracy from 16% (in paper) to 21%.

We considered this accuracy to be good because we also need to consider the fact that an image might belong to more than one genres. Therefore, our network might learn some features to classify into a certain genre, but a human might have used some other features to annotate it into some other category.

RESULTS FOR CLASSFICATION BASED ON ERA

We measured our performance using the accuracy of our classifier.

Accuracy = No. of correctly classified image/ Total number of images * 100

$$\text{Accuracy } = \frac{\text{No. of correctly classified image}}{\text{Total number of images}} * \ 100$$

We obtained an accuracy of ~50%. This is because our Classifier just used the image pixel for classifying them into 4 different eras.

CONCLUSION

We studied the evolution of website designs using a conjunction of various low-level features. We also classified any given images into either of the 4 eras and genres.

Our analysis about the textual information on a websites hints that in the earliest era (~1996-2000), the websites had very less content on their webpages. In the medieval period (~2000-2006), any given website had highest number of words. This is because, the websites were highly cluttered with words and disorganized. In terms of genres, educational websites had highest number of words. However, in the recent years, the word counts have reduced again. This implies that the websites have progressed towards better modularity and organized structure.

In case of color distributions using K-means & our colorfulness metric, we see that initially the websites were highly dull colored and were blank. As the internet evolved, the website designs used highly bold and bright colors. However, in the recent eras after 2010, as we saw in textual information, the color distribution has become moderate and have started using sophisticated colors that are aesthetically soothing.

Finally, our implementation of CNN for classifying images into various genres has improved accuracy from 16% to 21%. The paper used 5 layered CNN(AlexNet) for classifying images into eras and achieved an accuracy of ~55%. However, our 2 layered CNN gave an accuracy of ~50%.

INDIVIDUAL TASKS

We had an amazing time working on this project.
We divided our tasks as follows:

1. Data preparation to extract various features: Apurva
2. Extracting textual data from web pages and giving word count per image: Shyam
3. Extracting insights from colorfulness from images: Apurva
4. Finding dominant colors in a webpage and creating visualizations: Apurva
5. Implementing convolutional neural network for genre classification: Shyam
6. Implementing convolutional neural network for eras classification: Surbhi
7. Creating interactive visualizations in Tableau: Surbhi

FUTURE WORK

1. The data available to us is from 1996 to 2013. We aim at collecting recent data to look at recent trend changes.
2. We aim at considering other important low-level features like complexity of web pages, symmetry of images etc. to study the visual trends.

ACKNOWLEDGMENTS

We are extremely grateful to Prof David J. Crandall for providing us guidance and support to work on his research. We would also like to Thank Bardia Doosti, PHD student for providing us all the data files and details about his work on evolution of website designs. We take this opportunity to express our profound gratitude and deep regards to Andreas Bueckle for his exemplary guidance and constant encouragement throughout the course.

REFERENCES

Doosti, B., Crandall, D.J., & Su, N.M. (2017). A Deep Study into the History of Web Design. In Proceedings of the 2017 ACM on Web Science Conference. Troy: New York. (pp. 329-338). DOI: 10.1145/3091478.3091503

Chen, W., Crandall, D.J., & Su, N.M. (2017). Understanding the aesthetic evolution of websites: Towards a notion of design periods. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. Denver, CO. (pp. 5976-5987). DOI: 10.1145/3025453.3025607

http://homes.soic.indiana.edu/bdoosti/Temp/links.html

David Hasler, Sabine Susstrunk.(2003). Measuring colourfulness in natural images. Audiovisual Communication Lab(LCAV), Swiss Fed. Inst. Of Techn(EPFL),Lausanne, Switzerland

Katharina Reinecke and Krzysztof Z. Gajos. 2014. Quantifying Visual Preferences Around the World. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’14). ACM, New York, NY, USA, 11–20. DOI:http://dx.doi.org/10.1145/2556288.2557052