ANALYZING IMDB MOVIE POSTERS

Joonkyu Park
KAIST
joonkyu4220@kaist.ac.kr

ABSTRACT

This report covers the explorations and discoveries I had with thousands of movie posters as a term project of 2021 fall GCT675 - CulturoInformatics: Theory and Applications. From IMDb(Internet Movie Database), I scraped the posters of the top thousand movies each genre, sorted by popularity, according to IMDb's classifications. With this mass data, I ran pose estimation modules and face recognition modules to extract how the actors and actresses are generally placed in movie posters. I checked the number of people in each genre movie posters, what kind of emotions they had on their faces, the sizes of the faces, and the location distributions of the faces. I checked the distribution of the color usage in each genre by putting the pixel values in 3-dimensional RGB, HSV, and YUV spaces, then into 16 by 16 by 16 bins. I visualized and numerically compared the distribution of the pixel colors.

1. INTRODUCTION

Movie posters are elaborately designed by artists to attract the audiences. There are various poster designs, as diverse as the movies themselves. Some posters appeal the good looking actors featuring in the film, some show off the vivid, characteristic color-palettes used in the film, and some hide the information of the film to intrigue more people. However different they may be, we recognize a film poster right at the moment we see it, as easily as we recognize a human being even when there are billions of versions of us. Speaking of people recognizing other people, we also have stereotypes on genre movie posters, just like we secretly do with, say, people with facial hair or tattoos. For example, we expect two main characters joyfully looking at each other with bright background in romantic film posters, and a huge surprised face in the center with a lot of dark and groom color background in horror film posters. This is because movie posters, not like fine art, share the same purpose, which is to increase ticket sales. To achieve the goal, to entice as many audience as possible, movie posters must follow certain principles. If a poster designer does not follow these principles, it is likely that people will fail to understand what the film is about, or would not even recognize it as a movie poster in the first place. The stereotypical layouts work as grammatical baseline for both the designers and the audiences. As if they are sharing the same language, designers lean on former successful film designs to reach audiences, and the cinema-goers rightfully guess the genre and mood, or even the narrative, of the movie, based on the rules.

Indeed, there are codified thumb-rules in genre movie poster designs, but these rules doesn't tell anything about the distribution of the quantitative data in the real world. How many romance films have two people in their posters? What is the average size of the faces in horror film posters? In comedy films, are there happily smiling faces, or just funny faces which are not necessarily happy faces? Do romance films use more vivid colors than thrillers? The explorations and discoveries covered in this report start from asking how much information we could read out from simple quantitative features in the movie posters. I mainly focused on verifying or disproving the stereotypes that we have in genre movie posters, and some of the experiments are done to see the trend by time, or user ratings.

I scraped the posters of feature films from IMDb, 1,000 posters from each genre. The genres are comedy, sci-fi, horror, romance, action, thriller, drama, mystery, crime, animation, adventure, and fantasy, twelve in total. The top 1,000 movies are chosen by popularity, sorted by IMDb. Along with the poster, I made use of the released year, user scores, and content ratings. To narrow down the scope of interest into simple features, I counted the number of people in each poster, their sizes, emotions, and positional distributions. On the color usage, I put the pixel values in 3-dimensional RGB, HSV, and YUV spaces, then into 16 by 16 by 16 bins. According to my analysis, there were some genres showing significantly different number of people, size, and emotions, but the positional distributions and color usage distributions did not show genre-specific feature.

2. RELATED WORKS

There have been academic studies to characterize the personalities of graphic designs or to classify genre with neural networks.

Zhao et al. [6] trained the network with large datasets of image patches from web scraping, tagged with their characteristics, such as cute, energetic, creative, etc. Taking advantage of the deep ranking framework along with the convolutional neural networks (CNN) they successfully trained the network to identify the personalities of the input graphic design. Moreover, the trained network can crop an image to a given shape, or even transfer the style of the image, to best fit certain characteristics.

Wi et al. [5] also used CNN network to classify the genre of the given movie poster. They exploited the Gram matrix which is supposed by Gatys et al. [3] to represent the style of the image.

Besides these researches, there are also personal projects with deep neural networks predicting movie information based only on the given poster. Jing [4] used CNN models to predict genre, and reached about 65% top-3 accuracy. Another CNN model is trained to predict whether or not the movie made more revenue than budget, and the accuracy was around 70%.

These works are impressive as they are, but my project is not to be compared with theirs by performance or accuracy, since my purpose is to see the overall distribution of the features extracted from the genre movie posters. The goal here is to find out the genre-specific features that are understandable on the human level, not to predict the information the individual movie with the poster using deep learning networks.

3. METHOD

3.1 IMDb Scraping

I used a Python library Beautiful Soup [1] to scrape movie posters from IMDb. I collected 1,000 data from each IMDb genre classification, which are comedy, sci-fi, horror, romance, action, thriller, drama, mystery, crime, animation, adventure, and fantasy. The 1,000 movies are selected to be the top popularity movies, also sorted by IMDb. Expecting to see a trend by released year or user scores, I scraped the basic information of the movie along with the posters.

3.2 Face Recognition

Counting the number of people in the poster and looking at the positional distribution of them, face recognition modules worked relatively well, compared to pose estimation modules. Most of the pose estimation modules failed to locate people in the posters, probably due to the gap between the real-world images on which the module is trained and the artist designed movie posters. I used a Python module PAZ [2] to detect human faces in the movie posters. To make a consistent and fair analysis on every poster, I only handled the posters with height-to-width ratio above 1.38 and below 1.58, and reshaped the image into (1500, 1013) I saved the 2-dimensional coordinates of the bounding box with four numbers, two of them for the upper left corner and the others for the lower right corner. With the coordinates I saved the seven confidence score of the predicted emotions, ranging from 0 to 1, which can be angry, disgust, fear, happy, sad, surprise, or neutral.

3.3 Color Space

Having the resized, reshaped images, I put the pixel values, either in RGB, HSV, or YUV, into 3-dimensional bins, 16 by 16 by 16.

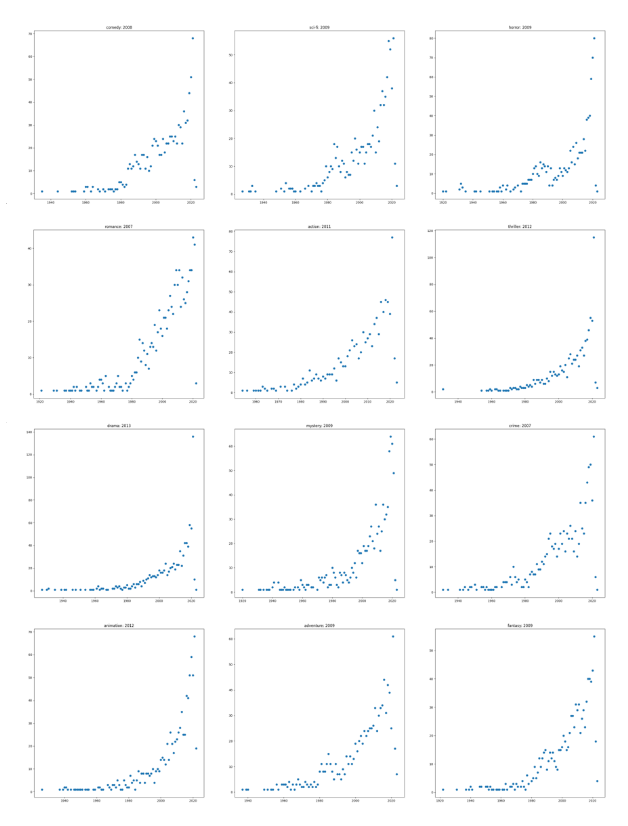

Figure 1. The distribution of the films by year, out of 1,000 feature films each genre with the highest popularity in IMDb.

4. RESULT

4.1 IMDb Scraping

To start with the IMDb scraping result, Figure 1 shows the distribution of the 1,000 most popular films in IMDb, each genre, by year. IMDb users' having interest more to recent films than to old films is expected and comprehensible, but the difference in the curves tell a lot about the activeness of the film industry in each genre. Romance films had the slowest decrease of popularity to the 1980s, which means that they don't easily get old, people having an on-going interest, compared to the other genres. For example, if we compare the curve of romance genre to that of thriller genre, people are mostly interested only in recent films that are not more than 2 years old. Thriller and drama show similar curves, but considering the total number of feature films are respectively 49,377 and 215,294, we can see that drama films are way more actively produced than thrillers, or we can also say that most of the genre-tags in feature films include drama.

Many genres have a kink, or burst, in 1980s, as is most apparent in sci-fi, horror, adventure, and fantasy. This gives a clue of when the golden era in film industry was especially in these genres.

Genre	Average Number of People
Comedy	**1.83**
Sci-fi	0.906
Horror	**0.672**
Romance	1.29
Action	1.27
Thriller	0.946
Drama	1.10
Mystery	0.756
Crime	1.19
Animation	0.434
Adventure	1.23
Fantasy	1.19

Table 1. Average number of human faces detected by PAZ face recognition module.

4.2 Face Recognition

The performance of PAZ face recognition module I used on movie posters is better than any other pose estimation modules but is still not 100% reliable. However, I expected the error to be negligible considering the amount of data I handle. Moving on to poster analysis, I first counted the average number of people in the poster, from each genre. As shown in Table 4.2, the number of faces detected in comedy films were significantly larger than in any other genres. Romance films had the second most detected faces, but still falls 29.5% short of that of comedy films.

Animation films has the least number of detected faces, but has no meaning, because it is merely an indication of how face recognition failed to detect the animated, probably even non-human, characters. Besides animation films, horror film has the least number of detected faces. Comparing comedy and horror, comedy films has 2.72 times more people(faces) in the poster, suggesting that ”the more, the merrier” works in genre films.

There is a different story when I looked into the sizes of the faces. Using the coordinate information of the upper left corner and the lower right corner of the face bounding boxes, I calculated the average size (in square-pixel units) of one face. Table 4.2 shows the result after extracting the square root of the average size. Mystery films have the largest average face size followed by horror films, where a comedy films have the smallest. Note that we can again neglect the result with animation films, since we cannot expect good accuracy from the face recognition module with the animated characters. When it comes to movie posters, Charlie Chaplin quote “life is a tragedy when seen in close-up, but a comedy in long-shot.” fits very well.

We can see from Table 4.2 and Table 4.2 that mystery/horror movies have few but large faces in their posters, where comedy films have many but small faces. What happens if we add up the sizes of the bounding boxes? Table 4.2 shows the averaged total area taken by face bounding boxes. Movie genre that devotes the largest area to showing the actors’ faces is neither comedy or horror, but romance. Romance films overcome the number difference with comedy films, and the size difference with horror films. It might give a romantic feeling showing close-up image of two people close to each other. Or, we might suspect that the romance films cast the most good-looking actors, so proud to appeal the cast in the front. On the other hand, sci-fi films used the smallest area in showing the characters’ faces, suggesting that sci-fi posters are busy showing the hi-tech graphic scenes, not having enough space spared for actors.

Genre	Average Size of a Face
Comedy	**143**
Sci-fi	166
Horror	206
Romance	185
Action	167
Thriller	196
Drama	187
Mystery	**222**
Crime	180
Animation	135
Adventure	151
Fantasy	155

Table 2. Average size of a face detected by PAZ face recognition module. The units are in pixels, for they are the square roots of the average bounding box sizes.

Genre	Average Total area of Faces
Comedy	194
Sci-fi	**158**
Horror	169
Romance	**209**
Action	188
Thriller	191
Drama	196
Mystery	193
Crime	197
Animation	90.0
Adventure	167
Fantasy	169

Table 3. Average total area taken by faces in a poster. The units are in pixels, for they are the square roots of the average total areas of the bounding boxes.

When it comes to emotion prediction scores, there were few significant differences between the genres. The confidence scores are added up, and then normalized to have a total score of 1. As shown in Table 4.2, faces in romance film posters were the most happiest, and the least angry/sad. Meanwhile, the sci-fi films had the most angry faces, and thriller films had the most sad faces in their posters.

Lastly, I visualized the positional distribution of the faces. I again used the coordinate information of the face bounding boxes, to acquire the center position of the bounding box, then put them into 2-dimensional 30 by 20

Genre	Angry	Happy	Sad
Comedy	0.090	0.30	**0.12**
Sci-Fi	**0.14**	0.13	0.17
Horror	0.13	0.12	0.17
Romance	**0.080**	**0.32**	**0.12**
Action	0.13	0.10	0.18
Thriller	0.13	**0.097**	**0.19**
Drama	0.11	0.19	0.16
Mystery	0.12	0.12	0.18
Crime	0.12	0.13	0.18
Animation	0.11	0.30	0.13
Adventure	0.12	0.17	0.16
Fantasy	0.10	0.23	0.14

Table 4. Sum of confidence scores from face recognition module. Seven scores (disgust, fear, surprise and neutral are omitted) are normalized to have a total value of 1.

bins. We can see in Figure 2 that all the genres obviously prefers to put a face in the in the middle, slightly above the center. However, in comedy movie posters, where there are many small faces, the distribution is clearly wide spread than in horror films, where there are few large faces.

I also tried filtering the coordinate data by number of people or emotion, but there was no noteworthy difference observed between the genres.

4.3 Color Space

I put the pixel values in 3-dimensional color space, either in RGB, HSV and YUV. Each movie had their own characteristic color palettes, as shown in Figure 3. However, when the genre movies are all added up, the distribution was hardly distinguishable. I present only an exemplar total distribution from comedy genre in Figure 4.

To find out the numerical difference between the color distributions, I came up with a simple metric system to calculate how colorful the distribution is. I measured the distance from each RGB space bin to a virtual straight line connecting (0, 0, 0) (black) and (15, 15, 15) (white), which has all the gray-scale colors. Then I weighted these distances with usage frequencies. According to my measure, animation film posters are significantly more colorful than any other genre, and the mystery film posters were the least colorful.

To see the contemporary trend in color usage frequencies, I divided the films into half, one group of recently released films, and the other half of relatively old ones. The colorfulness measure of comedy film posters increased significantly, where that of mystery film posters decreased by a huge gap. The measure of other genres maintained to similar extent through the years. The specific figures are shown in Table 4.3.

5. CONCLUSION

From face recognition results to color space distributions, I explored the genre movie posters to find out statistical dif-

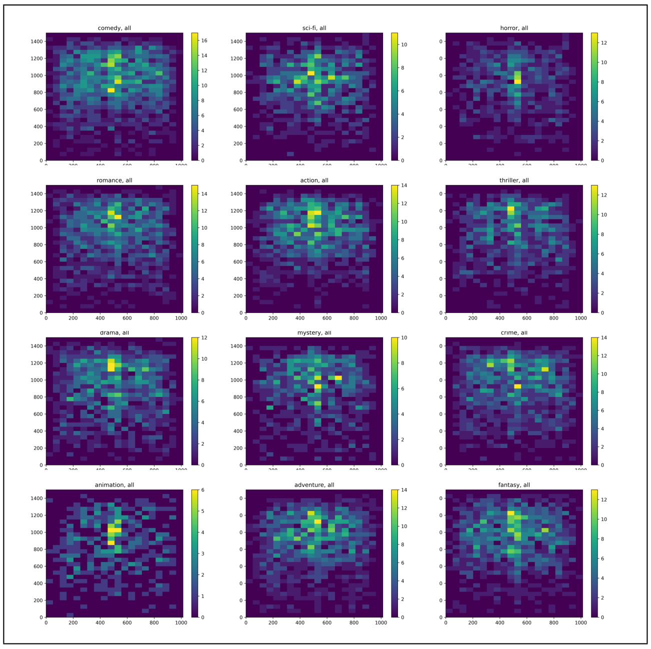

Figure 2. The positional distribution of the center of the detected faces. (1500, 1013) image is divided into (30, 20) bins.

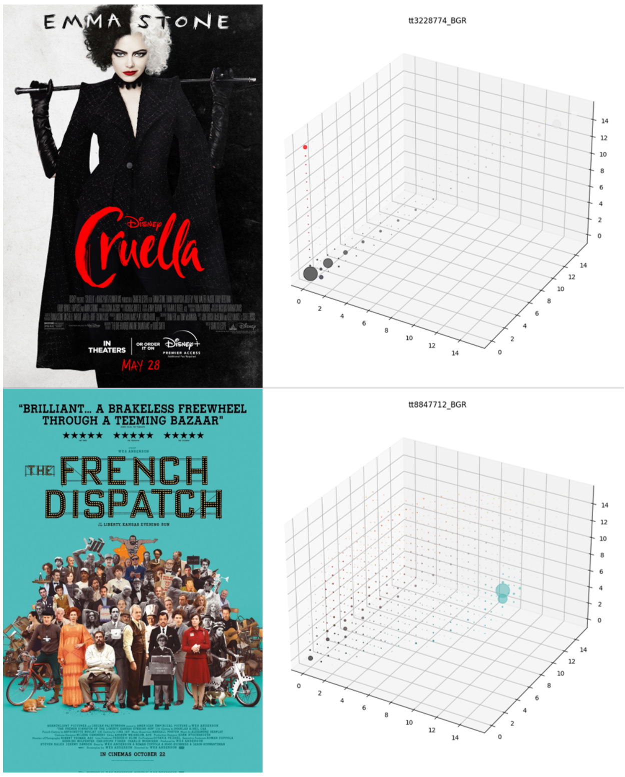

Figure 3. Each pixel value is distributed to 3-dimensional 16 by 16 by 16 bins in RGB space. The size of each circle is proportional to the usage frequency.

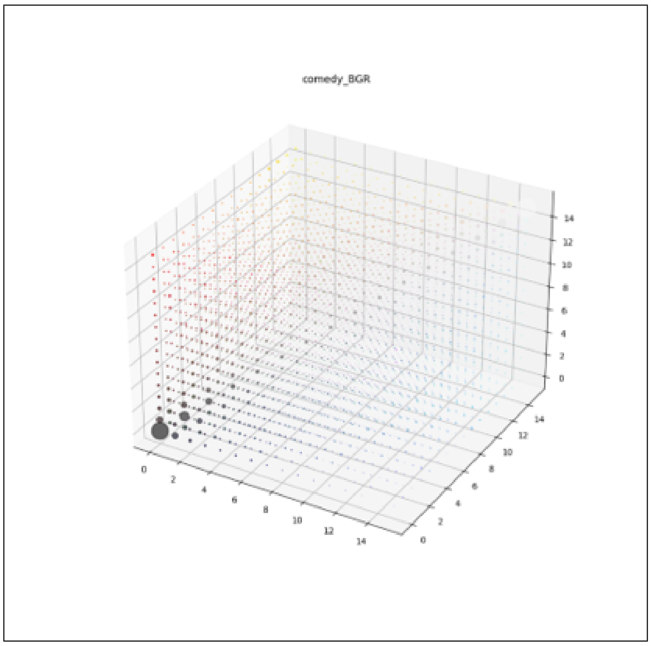

Figure 4. The total color distribution of comedy movie posters, visualized in RGB space.

Genre	Before	After
Comedy	**2.22**	**2.61**
Sci-Fi	2.09	1.97
Horror	1.78	1.72
Romance	2.20	2.14
Action	1.87	1.95
Thriller	1.71	1.76
Drama	1.83	1.89
Mystery	**1.81**	**1.61**
Crime	1.89	1.86
Animation	**2.87**	**2.96**
Adventure	2.30	2.25
Fantasy	2.22	2.28

Table 5. The colorfulness measure, indicating how much the pixel values are far from being gray-scale.

ferences between each genre. Comedy movies had many small faces in the posters, and mystery/horror movies had few large faces in the center. Romance movies spared the largest proportion in the poster to show the actors' faces, mostly happy faces. and thriller movies had the least happy faces. These discoveries are vulnerable to criticism on the performance of the face recognition module I used on poster images.

Animation films have the most colorful posters, and it got more colorful recently. On the other hand, mystery films have the least colorful posters, and it got even less colorful recently.

However interesting or expected the discoveries may be, there are still a lot to explore. Specifically, there are a lot to be done with individual films, not treating all the 1,000 genre movies as a whole. For example, there may be some correlation between a certain emotion, say, sad, and a certain color, say, blue or black. There may be some unexpected findings with the distribution of the distances between the detected faces. Categorizing the posters not by genre but by film company or marketing team, or a director, might also show a bias in choosing colors or in putting actors' faces.

6. REFERENCES

[1] Beautiful soup documentation.

[2] Octavio Arriaga, Matias Valdenegro-Toro, Mohandass Muthuraja, Sushma Devaramani, and Frank Kirchner. Perception for autonomous systems (paz), 2020.

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] Hong Jing. Predict movie earnings with posters, Mar 2020.

[5] Jeong A. Wi, Soojin Jang, and Youngbin Kim. Poster-based multiple movie genre classification using inter-channel features. *IEEE Access*, 8:66615–66624, 2020.

[6] Nanxuan Zhao, Ying Cao, and Rynson W.H. Lau. What characterizes personalities of graphic designs? *ACM Transactions on Graphics (Proc. of SIGGRAPH 2018)*, 37, 2018.