Decoding Eye-Catchiness: Exploring the Relationship Between Thumbnail Features and Viewer Metrics on YouTube

Lean Ting Jin
Matrikelnummer 6956985
ting-jin.lean@student.uni-tuebingen.de

Finn Springorum
Matrikelnummer 6124977
finn.springorum@student.uni-tuebingen.de

Christian Traxler
Matrikelnummer 6969273
christian.traxler@student.uni-tuebingen.de

Anna Chechenina
Matrikelnummer 6987499
anna.chechenina@student.uni-tuebingen.de

Abstract

It is widely believed that visually appealing YouTube thumbnails attract more viewers and thus generate more income for creators. In this study, we investigated this hypothesis by uncovering possible relationships between different features we associate with the eye-catchiness of a thumbnail and the view count, based on 80,000 entertainment videos collected with the YouTube Data API [1]. We normalized the view count based on the subscriber count and applied linear regression models to discover potential correlations. Our results indicate that videos with more views tend to be more eye-catching on average, but also underline the complex nature of this topic, as our thumbnail features were insufficient to explain variations in view count not accounted for by the subscriber count. To our knowledge, this is the first study to observe the effects of thumbnail visual appeal on video view counts.

1 Introduction

YouTube is the second most visited website in the world. In 2024, creators uploaded over 378 million hours of content to YouTube, and were paid more than 50 billion dollars in revenue (source: [2]). For YouTube creators, there is a massive financial incentive to maximize the views of their videos, whose promotion depends on the enigmatic YouTube algorithm. However, it is ultimately the user who decides in a split second whether a video is being watched. It is plausible that the thumbnail, a fundamental component of the video's presentation, plays a pivotal role in influencing user engagement and selection, as claimed by online resources [] and prominent YouTubers []. However, there does not seem to be any quantitative results on this topic in literature.

Admittedly, it is not clear how one can quantify the eye-catchiness of a thumbnail, but reasonable to assume that certain features such as image color, saturation, and the presence of human faces are important. We hypothesize that eye-catching thumbnails have a positive effect on the number of views the corresponding video receives. As a proxy for eye-catchiness, our study is based on six features which are derived from the thumbnail: hue, saturation, lightness, contrast, sharpness, and the number of faces.

2 Methods

Data Collection. We utilized the YouTube Data API [1] to collect data on videos from a variety of categories. There are 15 main categories for videos on YouTube, but given our interest in thumbnails, we decided to focus on videos intended to appear in the feed of a user and elicit a response. Therefore, the following eight categories were used: Comedy, Education, Entertainment, Gaming, How-to & Style, News & Politics, People & Blogs, and Sports. We excluded categories corresponding to videos explicitly sought after through searches (e.g., Film & Animation), videos which are unlikely to include human faces on their thumbnail (e.g., Pets & Animals), or videos from a very niche field (e.g., Nonprofits & Activism).

Due to the limited number of videos that can be retrieved for a given query, the date range from January 1st, 2015 to the time of collection was divided into disjoint sets, and 500 videos were requested for each period. This strategy enabled the acquisition of thousands of videos for a single query. To prevent duplicates while respecting the daily API limit, our data collection strategy entailed the request of up to 2,500 videos from a designated category, each accompanied by a singular generic keyword (e.g., most popular video games for the Gaming category), until the collection comprised precisely 10,000 unique videos from said category, yielding a dataset with eight categories, for a total of 80,000 unique videos.

Video and Thumbnail Features. For each video, we collected the thumbnail, view count, and subscriber count. From each thumbnail, we extracted the following 6 features: hue, saturation, lightness, contrast, sharpness, and the number of faces.

The hue is the angle of the mean color of the thumbnail in the HSV color space, with a range of 0 to 360 degrees. The saturation, lightness, and contrast follow their standard definitions as implemented in the OpenCV Python library (normalized to [0,1]). The sharpness is the log variance of the Laplacian of the grayscale image. The number of faces is the number of distinct faces in the thumbnail as predicted by the RetinaFace model implemented in the DeepFace library [3, 4].

Analysis. We are interested the the relationship between the view count N_i and the six features derived from the thumbnail T_i. However, as implied by intuition and a linear regression analysis (Fig. 1), there is a significant positive correlation between the view count of a video and the number of subscribers S_i of its corresponding YouTube channel. To remove this effect and focus on variations in view count not explained by the subscriber count, we normalized the view count by applying linear regression (LR) models of the log view count against the log subscriber count for each category (log: base 10). The log-transformation ensures that the residuals are approximately homoscedastic and normally distributed, as the view and subscriber count distribution appears to be log-normal. This way, we obtained the residuals $R_i := \log(N_i) - \hat{\beta}_0^{(\gamma)} - \hat{\beta}_1^{(\gamma)} \log(S_i)$, with $\gamma \in \{1, ..., 8\}$ corresponding to one of the eight categories that video V_i belongs to. We refer to these residuals as "residual log view count" and use them as well as the non-normalized "log view count" for all models in the following analysis.

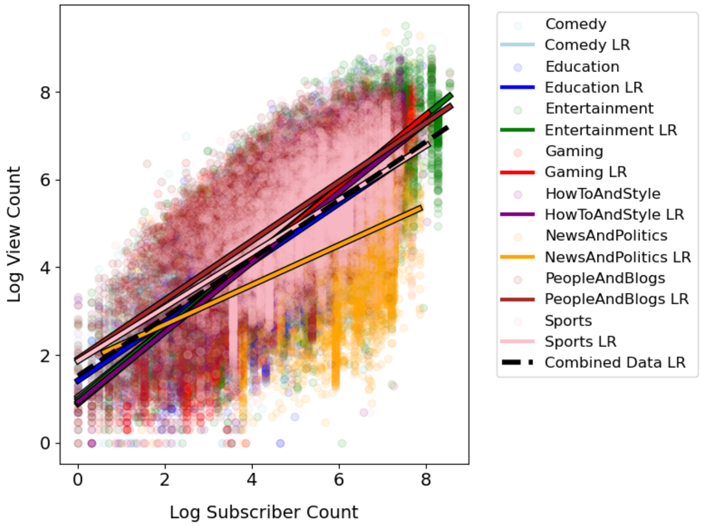

Figure 1: Linear regression (LR) of the log view count against the log subscriber count. The relationships resemble for all eight video categories. For the combined dataset, the relationship is $\widehat{\log(N_i)} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(S_i)$, 95% CI for $\beta_0 = [3.445, 3.571]$, 95% CI for $\beta_1 = [0.663, 0.673]$, $R^2 = 0.467$.

Furthermore, the correlation between the six thumbnail features was investigated. There did not appear to be strong multicollinearity between features, except for the lightness-contrast correlation which was highly positive ($c = 0.79$). A random forest regression model with 500 trees and three of six features considered per split was applied to obtain some non-linear insights about the feature importance.

Moreover, a linear regression model was fitted to each continuous feature (saturation, lightness, contrast, sharpness). Meanwhile, the number of faces and the hue were treated as categorical features, and 95% confidence intervals of the log view count and residual log view count were computed for each category. The face count was divided into the categories 0, 1, 2, and 3+, with the last category containing all videos with at least three detected faces on their thumbnail. The hue was divided into six bins corresponding to a 60-degree range.

Finally, the residual log view count was fitted against saturation, contrast, sharpness, and the number of faces in a multiple linear regression model. We removed the lightness feature due to the multicollinearity mentioned above, and the hue due to its unique and non-linear characteristics.

3 Results

We focus on the entire dataset with nearly 80,000 valid data points here, since we could not observe significant differences between the category datasets. For both the non-normalized log view count and normalized residual log view count, the feature importance results of the random forest model were similar for the five simple image features hue, saturation, lightness, contrast, and sharpness (around 0.19), with slight variations between the categories, whereas the number of faces yielded a smaller importance of about 0.05.

The linear regression models for the continuous features (Fig. 2) reveal a positive slope for the log view count against each feature, respectively, a smaller positive slope for the residual log view count against saturation, and even a negative slope for the relationship between normalized views and lightness, contrast, and sharpness, respectively. The 95% confidence intervals for the face count and hue (Fig. 3) also exhibit major differences between the non-normalized and normalized view counts: Whereas the log view count intervals for thumbnails with at least one face lie higher than and do not overlap with the interval for no faces, one cannot observe this trend for the residual log view count. For the hue, however, the confidence intervals for both types of view counts with a mid-range average hue lie clearly above those of the 0-60 and 300-360 degree range.

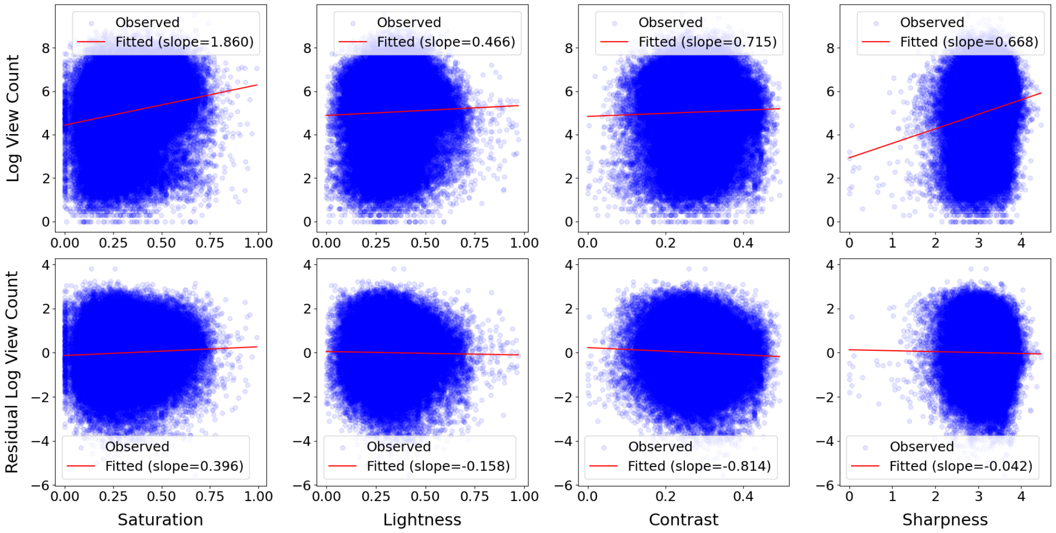

Figure 2: Linear regression results for the log view count (top) and the residual log view count (bottom) against the four features saturation, lightness, contrast, and sharpness, respectively, based on the entire dataset.

For the multiple linear regression model based on the four features saturation, contrast, sharpness, and number of faces, we obtained p-values negligibly close to zero for the overall model. For the non-normalized view count, all p-values for the different features were zero, and all coefficients

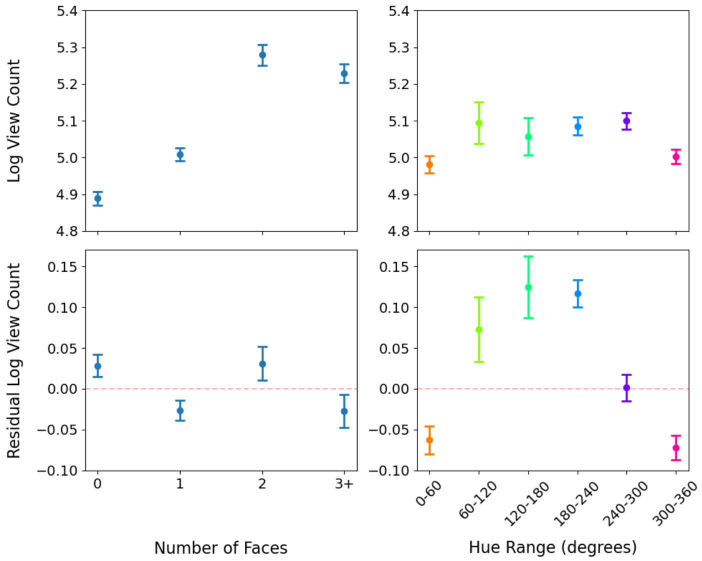

Figure 3: 95% confidence intervals of the log view count and residual log view count for each face and hue category, as previously defined. The color of each hue boxplot was chosen according to the central hue of the corresponding bin. Number of samples per view count: [0: 26,850; 1: 28,696; 2: 10,872; 3: 12,577]. Number of samples per hue bin: [0-60: 16,530; 60-120: 3,079; 120-180: 3,681; 180-240: 16,243; 240-300: 17,370; 300-360: 21,826].

except for the contrast coefficient were positive (saturation: 1.55, contrast: -1.41, sharpness: 0.67, num_faces: 0.02). For the normalized view count, the p-value for sharpness was 0.294, the others were zero as well, and the coefficients were mostly negative or close to zero, except for the saturation coefficient (saturation: 0.42, contrast: -0.91, sharpness: 0.01, num_faces: -0.01).

4 Discussion/Limitations

Findings. The near-zero p-values of our multiple linear regression results indicate a significant relationship between at least one of the features we associate with the eye-catchiness of a thumbnail and the view count of the corresponding video. However, one must distinguish between the non-normalized and the normalized version of the view count. For the former one, we observed positive slopes in the single regression models, and a significant difference between videos with and without faces on their thumbnails. Although these observations support our hypothesis, the causality remains unclear, since these findings could also imply that more successful channels produce more eye-catching thumbnails, but obtain more views because of other features not related to our eye-catchiness. Most importantly, the clear correlation with the subscriber count must be addressed. After normalizing the view count, our regression plots no longer showed positive relationships. This suggests that the eye-catchiness, as defined by us, is likely insufficient to explain the variations in view count for a fixed number of subscribers. Only the saturation feature coefficients were always positive, which may indicate that saturation is the most important feature to optimize as a YouTube creator. The random forest model did not provide valuable information either, since the five simple image features were relatively similar in terms of predictive power, only the face count had a much lower feature importance which might be due to its discrete and thus limited nature. The hue, which was excluded for the linear regression methods, showed an interesting level of consistency across non-normalized and normalized view counts, indicating that thumbnails with predominantly green and blue colors might yield more views on average as opposed to red ones. However, this could also be due to the small sample sizes for some of these bins.

Other limitations. We found linear regression models most plausible for our study, since more complex relationships are highly unlikely. However, these initial assumptions might not hold, and perhaps it is not even possible to find empirical support for our hypothesis, since the YouTube algorithm might promote videos without highly eye-catching thumbnails so that many users click on them anyway. Furthermore, inherent limitations of the YouTube API need to be considered, for instance the overlapping and vague video categories that are either user-defined or automatically assigned, leading to the unwanted inclusion or exclusion of samples. In addition, despite our dataset being reasonably large, it cannot represent the vast landscape of YouTube videos, since the algorithm decided which videos we were able to collect, and the majority of them had views in the five-figure range or higher.

Text detection.

5 Statement of Contributions

Chase and Finn set up the API scripts, performed the correlation and regression analysis, and created the plots. Christian wrote API scripts to collect the subscriber counts for our dataset. Anna implemented and analyzed the text detection methods and models. All members of the group contributed to collecting the dataset and writing the report.

References

[1] Google. YouTube Data API v3. https://developers.google.com/youtube/v3/docs.

[2] https://www.marketingscoop.com/marketing/how-many-hours-of-video-are-uploaded-to-youtube-ever

[3] Sefik Ilkin Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Bilisim Teknolojileri Dergisi*, 17(2):95–107, 2024.

[4] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE, 2020.