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

Lean Ting Jin
Matrikelnummer 6956985

Finn Springorum
Matrikelnummer 6124977

Christian Traxler
Matrikelnummer 6969273

Anna Chechenina
Matrikelnummer 6987499

Abstract

Visually appealing YouTube thumbnails are believed to increase viewership and thus generate more income for creators. In this study, we investigate this hypothesis by analyzing the relationship between features associated with thumbnail eye-catchiness and the view count, based on 80,000 entertainment videos collected with the YouTube Data API [1]. We normalized view counts relative to subscriber counts and applied linear regression models to identify potential correlations. Our results suggest that videos with higher view counts tend to have more eye-catching thumbnails. However, thumbnail features were insufficient to explain variations in view counts unaccounted for by subscriber counts. To our knowledge, this is the first study to examine the impact of thumbnail visual appeal on 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 [2]. For YouTube creators, there is a significant 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 watched. Thumbnails, the primary visual element presented to users, appear to play a crucial role in influencing user engagement and selection, as suggested by YouTube's official resources [3]. However, there appears to be no quantitative research on this topic in literature.

Quantifying the eye-catchiness of a thumbnail is inherently challenging. However, certain features, such as image color, saturation, and the presence of human faces, likely contribute to visual prominence. As a proxy for eye-catchiness, our study examines six thumbnail-derived features: hue, saturation, lightness, contrast, sharpness, and the number of faces. We hypothesize that more eye-catching thumbnails lead to higher video view counts.

2 Methods

Data Collection. We used the YouTube Data API [1] to collect video data from eight of YouTube's 15 categories: Comedy, Education, Entertainment, Gaming, How-to & Style, News & Politics, People & Blogs, and Sports. These videos are primarily designed to appear in a user's feed and elicit a response, making them suitable for our thumbnail study.

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 intervals, and 500 videos were requested for each interval. To prevent duplicates while respecting the daily API limit, we collected 2,500 videos from a designated category using a distinct 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 a total of 80,000 unique videos.

Video and Thumbnail Features. For each video, we collected its thumbnail, view count, and subscriber count. From each thumbnail, we extracted six features: hue, saturation, lightness, contrast, sharpness, and the number of faces.. Hue, saturation, and lightness represent the average values of the thumbnail in the HSL image format [4] (as implemented in OpenCV [5]), where hue ranges from 0 to 360 degrees, and saturation and lightness are normalized to [0,1]. Contrast is defined as the root mean square contrast, which is the standard deviation of the grayscale image [6]. Sharpness is measure as the log variance of the Laplacian of the grayscale image [7]. The number of faces is the number of distinct faces in the thumbnail as predicted by the RetinaFace model implemented in the DeepFace library [8, 9].

Analysis. We were interested the the relationship between the view count N_i and the six features derived from the thumbnail T_i. However, intuition expects correlation between a video's view count and a channel's subscriber count, S_i, and a linear regression (LR) model (Fig. 1) confirms this relationship with a strong positive Pearson correlation ($\rho = 0.68$). To isolate variations in view count independent of subscriber count, we applied a linear regression model, regressing log-transformed (base 10) view count against the log-transformed (base 10) subscriber count for each category. 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. We obtained the residuals $R_i := \log(N_i) - \hat{\beta}_0^{(\gamma)} - \hat{\beta}_1^{(\gamma)} \log(S_i)$, where $\gamma \in \{1, ..., 8\}$ corresponds the category of video V_i. We refer to these residuals as the "residual log view count" and use them, along with the non-normalized "log view count" in the following analysis.

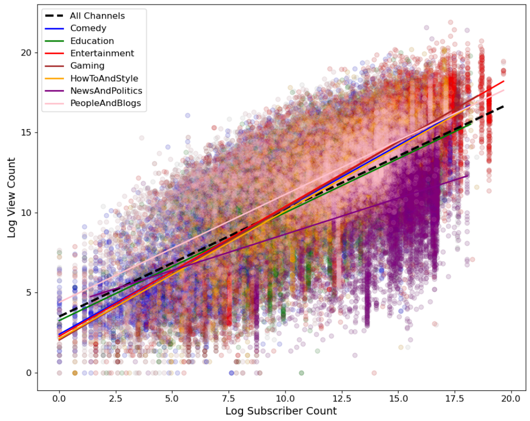

Figure 1: Linear regression (LR) of the log view count ($\log(N_i)$) against the log subscriber count ($\log(S_i)$) across the eight video categories. The relationships are consistent across categories. For the combined dataset, the relationship is given: $\widehat{\log(N_i)} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(S_i)$, with 95% confidence intervals (CI) for β_0 as $[3.445, 3.571]$, for β_1 as $[0.663, 0.673]$, and $R^2 = 0.467$.

Furthermore, the correlation between the six thumbnail features were analyzed. No strong multicollinearity between features was observed, except for a strong positive correlation between lightness and contrast ($\rho = 0.79$).

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, with 95% confidence intervals (CIs) of the log view count and residual log view count computed for each category. The face count data was categorized into four groups: 0, 1, 2, and 3+ (the last category comprising 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, both view counts were fitted against saturation, contrast, sharpness, and the number of faces in a multiple linear regression model. We omitted the lightness feature due to the multicollinearity mentioned above, and the hue feature due to its unique and non-linear characteristics.

3 Results

We focus our analysis on the entire dataset with nearly 80,000 valid data points, since we did not observe significant differences between the eight video categories.

The linear regression models for the continuous features (Fig. 2) revealed a positive slope for the log view count against each feature individually, 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 exhibited major differences between the log and residual log view counts. While the log view count intervals for thumbnails with at least one face was higher and did not overlap with the interval for no faces, this trend was absent 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 lay clearly above those of the 0-60 and 300-360 degree range.

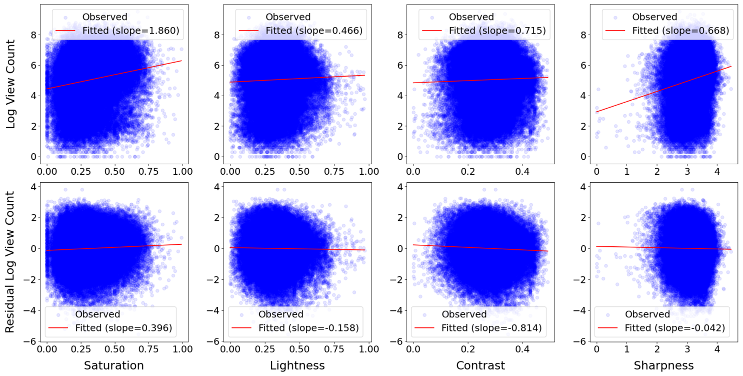

Figure 2: Linear regression results for the log view count (top) and the residual log view count (bottom) against the four features: saturation, lightness, contrast, and sharpness, respectively, across the entire dataset.

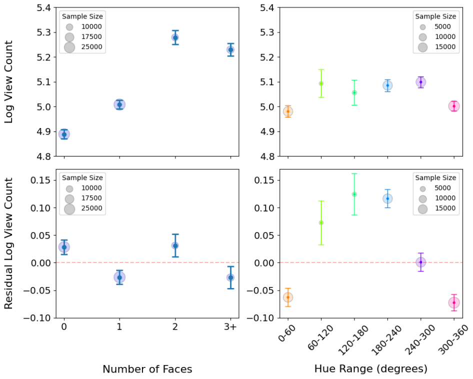

Figure 3: 95% CIs of the log view count (top) and residual log view count (bottom) for different face and hue categories. The hue boxplot colors correspond to the central hue of each bin, and marker sizes are proportional to the sample size.

	Non-normalized		Normalized	
Feature	95% CI	p-value	95% CI	p-value
Saturation	[1.482, 1.627]	$< 10^{-308}$	[0.362, 0.470]	$< 10^{-308}$
Contrast	[-1.587, -1.236]	$< 10^{-308}$	[-1.043, -0.780]	$< 10^{-41}$
Sharpness	[0.637, 0.699]	$< 10^{-308}$	[-0.011, 0.036]	0.294
Number of faces	[0.012, 0.020]	$< 10^{-13}$	[-0.013, -0.006]	$< 10^{-8}$

Table 1: 95% coefficient CI and p-value for each feature in the two multiple linear regression models. $R^2 = 0.055$ (log view count) and $R^2 = 0.006$ (residual log view count).

For the multiple linear regression model (Tab. 1) 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 almost zero, and all coefficients except for the contrast coefficient were positive. For the normalized view count, the p-value for sharpness was 0.294, the others were almost zero as well, and the coefficients were mostly negative or close to zero, except for the saturation coefficient.

4 Discussion/Limitations

Findings. The near-zero p-values of our multiple linear regression results indicate a significant relationship between multiple 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. While 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 due to other features unrelated to eye-catchiness. Most importantly, the clear correlation with the subscriber count must be addressed. After normalizing the view count, most of 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. Interestingly, the analysis for the hue was consistent across non-normalized and normalized view counts, indicating that thumbnails with predominantly green and blue colors might yield more views on average compared to reddish ones. Although, this must be further investigated as it may be due to the HSL [4] representation where red colors can correspond to hues of 0 degrees or 360 degrees. Meaning that predominantly red thumbnails may average out to 180 degrees, which is in between blue and green. Future work could investigate the difference between using different color representations such as RGB, or even finding different metrics for color instead of averaging values such as color quantization [**?**]. Notably, applying a simple random forest model did not yield meaningful results.

Other Limitations. We found linear regression models to be the most plausible for our study, as more complex relationships are highly unlikely. However, these assumptions might not hold, and the linear regression was inaccurate and distorted by outliers, particularly in the sharpness regressions. Moreover, there are some inherent limitations of the YouTube Data API. Firstly, the video categories are somewhat overlapping and vague, being user-defined or automatically generated. Seeing how very little difference between categories was found during our analysis, future work could choose to ignore categories. Additionally, the API does not grant access important data, such as number of impressions, average view duration, or click-through-rate for video which could all be important factors to consider. Since this data is not public, future work could partner with big channels to perform analysis on their data. Furthermore, despite our dataset being reasonably large, it cannot be fully representative of all YouTube videos, as the video search algorithm favors popular and recent videos.

Text Detection. Initially, we intended to consider the presence of text on the thumbnails as an additional eye-catching feature. However, during the analysis and implementation, we observed poor accuracy of the models, even after attempting to identify the language from the video titles. Consequently, we decided not to include this feature in our study and leave it open for future research. Future work could include such analysis by only analyzing videos in languages employing the Roman alphabet to decrease the error rate in the text detection models.

5 Statement of Contributions

Chase and Finn set up the main 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] Michael Smith. How Many Hours of Video Are Uploaded to YouTube Every Minute? https://www.marketingscoop.com/marketing/how-many-hours-of-video-are-uploaded-to-youtube-every-minute/, Nov 2024.

[3] YouTube. Thumbnail and title tips. https://support.google.com/youtube/answer/12340300.

[4] Max K. Agoston. *Computer Graphics and Geometric Modeling*. Springer, 2005.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] Eli Peli. Contrast in complex images. *Journal of the Optical Society of America*, 7(10):2032, Oct 1990.

[7] J.L. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia. Diatom autofocusing in brightfield microscopy: a comparative study. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 314–317 vol.3, 2000.

[8] 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.

[9] 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.