Detecting Cultural Differences in News Video Thumbnails via Computational Aesthetics

Marvin Limpijankit
Graduate, Computer Science Dept.
Columbia University, NY 10027
ml4431@columbia.edu

John R. Kender
Advisor, Computer Science Dept.
Columbia University, NY 10027
jrk@cs.columbia.edu

Abstract

Images published by news sources may be indicators of cultural differences as the selection of which images to use (content) and how they are edited (style) may reflect the media's intentions as well as the interest of their respective audiences. Many studies have qualitatively assessed how news images emphasize certain narratives, however, these are limited in scale and the ability to quantify results. Alternatively, purely computational comparisons of statistical properties struggle with the fact that these properties are highly dependent on the actual content of the image itself. To address these issues, we propose a two-step approach for detecting differences in images across sources of differing cultural affinity, where images are first clustered into finer sub-topics based on content before their features are compared. We test this approach on a set of YouTube thumbnails taken from Chinese and US YouTube channels relating to COVID-19 and the Ukraine conflict. Our results suggest that while Chinese thumbnails are less formal and more candid, US channels tend to use more deliberate, proper photographs as thumbnails.

Index Terms: news media, cross-cultural analysis, computational aesthetics

1 Introduction

The widespread adoption of technology in recent times has contributed greatly to a shift away from traditional news outlets to ones hosted on social media. From 2021 to 2022, U.S. Daily newspaper circulation (print and digital combined) has decreased by 8%, and approximately half of Americans now indicate that they consume news from social media platforms, with Facebook and YouTube being among the most popular [1, 2]. As a result of this digitization, the online space has become an incredibly data-rich resource for identifying key differences in how culturally distinct sources choose to present global events.

Visual imagery is an extremely important channel for international news coverage as it connects the audience to events that may otherwise be removed from direct experience, while also serving as a powerful tool for conveying emotion [3]. For threatening events, such as global warming and COVID-19, images have also been shown to be significant drivers of public engagement, "providing viewers with tangible and emotion-evoking examples that act as visual proof" [4, 5]. As such, highly visual mediums such as YouTube, the largest video-sharing platform with over one million hours of news-related content uploaded weekly, play a significant role in modern news consumption [6, 7]. In this paper, we attempt to apply computational approaches to quantify any differences in the use of visual imagery between news sources from varying regional backgrounds. More specifically, our work lies at the interplay of sociology and computer science, adopting techniques from computer vision to extract visual features from YouTube thumbnails at scale, before using methods such as dimensionality reduction to accentuate cross-cultural differences between US and Chinese channels. We focus on two international events, the COVID-19 pandemic, and the war in Ukraine. Since statistical image properties are highly dependent on the content of the image [8, 9], as a preprocessing step, our proposed framework clusters images from a given news story by 'visual themes' (i.e. different aspects or sub-topics) before analyzing between channels. The clustering is analogous to

topic modeling in NLP. Finally, we also consider the views, likes, and comments for each video in order to capture any potential cross-cultural differences in how thumbnail properties influence video performance.

To summarize, in this paper we

- propose a framework to compare image properties in YouTube thumbnails while accounting for visual content.
- run this procedure on a dataset of 2400 images from 2 international events and 4 news channels.
- report how the visual properties of thumbnails and viewership statistics differ between US and Chinese sources

2 Related Work

Computational Aesthetics. In automatic image aesthetic assessment, also known as computational aesthetics, a model is trained to quantify the beauty of an image through some aesthetic score[1]. Often, these models assess an image by measuring the degree to which it adheres to fundamental principles of photography (e.g. balance, texture, unity) [10]. There are two main approaches to image aesthetic assessment, hand-crafted methods, and deep learning methods. Hand-crafted methods rely on extracting pre-defined low-level features from images and fitting a model, such as SVM, to do the prediction. Examples of features that have been extensively explored previously include sharpness, colorfulness, contrast, and texture [11, 12, 13]. Deep learning methods leverage neural networks to learn low and high-level image features without (or with relatively little) human instruction. These are trained on large image datasets such as Photo.net [14], AVA [15], and CUHK-PQ [16], and the specific architecture can vary. Want et al. fine-tuned two pre-trained models (AlexNet and VGG) to classify images into high and low aesthetic categories. Kao et al. investigated using a Multi-Task Convolutional Neutral network (MTCNN) to jointly perform aesthetic assessment and semantic recognition [8].

In addition to predicting image aesthetics (quality, beauty, liking) [17, 18], there is evidence to suggest that computational aesthetics can serve as a valuable starting point for a variety of downstream tasks. For instance, one experiment demonstrated the ability to infer the personality traits (BFI) of Flickr users based on the color, composition, texture, and face features of images the user tagged as 'favorites' on the platform [19]. The authors emphasize that this approach was only particularly effective for predicting self-attributed traits (i.e. the personality impressions that the images convey) and not necessarily self-assessed traits (i.e. the personality impressions the user identified for themselves) to the same degree. Machajdik et al. achieved state-of-the-art results on image emotion classification using a similar set of features, including color, texture, composition, and content, inspired by psychology and art theory [9]. They note that certain features outperform others depending on the dataset and emotion being predicted. Content features such as the number of faces and relative size of the biggest face were good discriminators between fear/disgust vs. amusement in the IAPS dataset whereas in the artistic photography dataset, the importance of color features was more prominent. The authors attribute this to the former dataset being more contextual (i.e. fear/disgust was portrayed through images of insects or injuries while those of amusement often contained portraits of people smiling) and artistic photography being more intentional in the use of color.

These studies reflect 1) the ability of image aesthetics to emphasize certain feelings and 2) the importance of considering the content within images when comparing aesthetic features.

Figure 1: "Photo Space" visualization [25]

Image Analysis in News Media. Cross-cultural differences in news portrayal have been studied extensively from both a sociological and computational perspective. The importance of visual narratives in mass media as a means of cultivating particular understandings of events is especially of interest. One investigation analyzed images of the COVID-19 pandemic appearing in Finnish newspapers, focusing on the social representations of individuals from varying age groups [5]. The authors noted that the angle of the photo often contributed to visual rhetoric, for instance, images of children studying at home emphasized a high angle, placing the spectator in the authoritative position overseeing the child

[1] Can also be formulated as a classification task

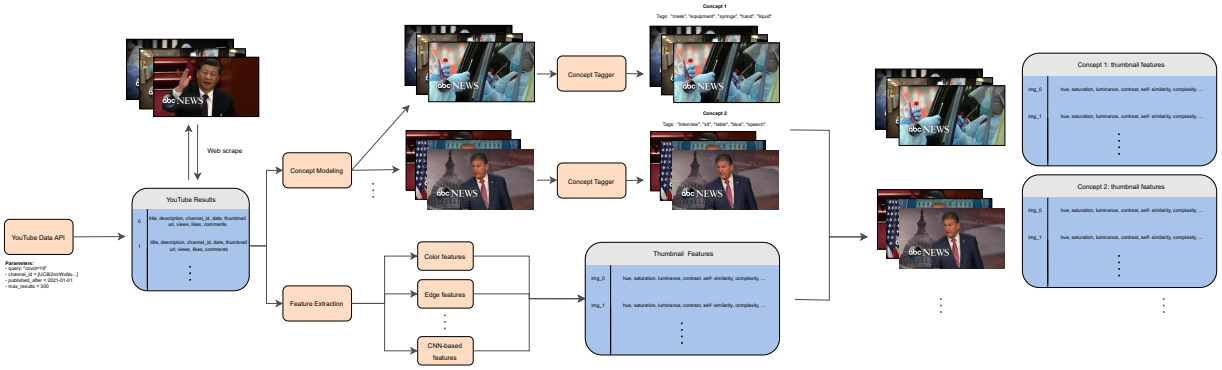

Figure 2: Our proposed framework. The URL for video thumbnails, along with metadata and viewership statistics, are scraped using the YouTube API given search parameters. A Python script downloads the thumbnails, then passes them to concept modeling to sort the images by visual themes. Tags are generated for each concept group. In parallel, feature extraction is performed. The results from both processes are then joined together.

whereas those of young adults tended to be taken at eye-level. Color also played a significant role, with red hues (with the connotation of sin) emphasizing the recklessness in images of young adults partying and darkness embodying the feeling of loneliness in pictures of children attending school online. Similarly, in a study on how the media covered Pope Francis' visit to Cuba, Thomson et al. discovered that photo angles in Western sources framed the Pope at higher positions than Cuban politicians whereas local sources used a more equal leveling [20]. The authors suggest this may reflect higher deference toward government officials in Cuban culture as opposed to placing religious figures above politicians. Another contrast was the content of the images, with Cuban media focusing almost exclusively on the Pope whereas Western media more often published photographs of Cuban people (onlooking the visit, holding posters, or protesting), demonstrating a difference in the visual narrative between the Pope's visit and the impact of his visit on everyday citizens. Additional investigations have also contrasted the visual discourse in media coverage of events [21, 22].

Computational approaches have been adopted to quantify differences and overcome issues surrounding the scalability of qualitative sociological analysis alone. Although these frequently involve analyzing text data to compare sentiments and topics [23, 24], the visual medium of news media has also been explored. Bhargava et al. proposed photo-spaces, a tool to describe visual narratives for an event based on image embeddings produced by a pre-trained ResNet-50 model [25]. The spatial positioning of images in the photo-space is determined by reducing these embeddings into 2 dimensions via UMAP, and from this clusters corresponding to visual themes can be extracted. The authors document an example of the tool (figure 1) on abortion rights as a topic, adding the political affinity of sources (right vs. left-leaning) as an extra layer in the visualization through the color of each image's border. Similarly, Zhang et al. evaluated the effectiveness of bag-of-visual-words, self-supervised learning, and transfer learning to cluster unlabeled image data [26]. They conclude that transfer learning, using pre-trained models to obtain embeddings, outperforms the other methods, though the choice of pre-trained model (namely which dataset the pre-training was done on) significantly affected clustering results.

In our investigation, we adopt image clustering as a pre-processing step in our pipeline, grouping images before examining cross-cultural differences as a way of controlling for visual context.

Video Thumbnails. Thumbnails are small image representations of full-length videos that users click on to be redirected to watch the video. Ideally, the role of the thumbnail should be to "tell the users what the video is about (i.e. be informative), and ... grab potential viewers' attention (i.e., be visually appealing)" [27]. Kim et al. evaluate the design of thumbnails for data-driven news articles, concluding similarly that for thumbnails involving informative graphics, users prefer ones that are interpretable and attractive [28]. For branded content, a thumbnail's success (measured in view-through) can also be attributed to visual features, such as colorfulness and brightness [27]. As such, thumbnails may serve as an interesting visual medium for cross-cultural news analysis, reflecting both the content the media wishes to present (information) as well as their attempts to appeal to specific cultural perspectives (attraction).

3 Methods

3.1 Data Collection

To evaluate cross-cultural differences, we select two US and two Chinese YouTube channels. Our criteria for selecting these news channels are that they 1) upload frequently enough to provide an adequate number of thumbnails per news event and 2) have minimal use of watermarking so as to decrease any potential effect on the extracted features. Channels with larger audiences are preferred due to having richer viewership data. The US sources are ABC and CBS, with 15.6M and 5.5M subscribers respectively, and the Chinese sources are CGTN (Chinese Global Television Network) and New China TV, with 3.0M and 1.4M subscribers respectively.

Inspired by previous work [5, 29], we select the COVID-19 pandemic as one of the international news events to examine. The international importance of the pandemic along with the length of the event lends itself well to this case study from a data magnitude perspective. Furthermore, the complexity of the event gives rise to a potentially diverse set of story angles to the event (e.g. medical, economic, political implications), which may transfer to detectable visual themes (concepts). For similar reasons, we also investigate the war in Ukraine, a more recent event that has gained large media attention and contains various story angles such as the military conflict, the humanitarian crisis, and geopolitics.

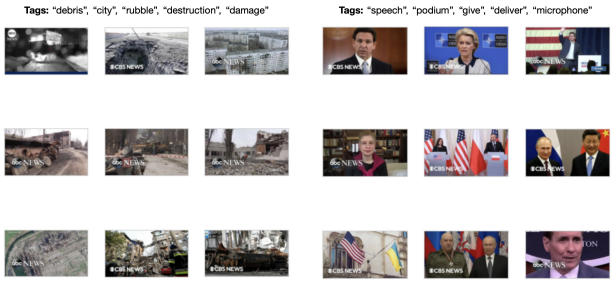

Figure 3: Example concepts from the Ukraine conflict

The YouTube API [30] was used to scrape the video metadata using a search query ("covid 19", "ukraine war"), the channel_ids of the above, and a published_after date to filter the videos to the appropriate event. For each event, channel, we collect the top 300 videos (sorted by YouTube's default 'relevance' parameter) as past this number, for at least one of the event, channel pairs, the results begin to appear unrelated to the queried event. Iterating through the results, image content is then scraped using a simple Python script and preprocessed to remove any vertical and horizontal black bars using the Pillow package. Each thumbnail becomes associated with an image_id with channel, views, likes, and comment data attached.

Event	Concept	Tagging Method		
		Approach 1 (top 5)	Approach 2 (top 5 adjusted)	Approach 3 (tf-idf)
COVID	0	woman, news, person, stand, man	interview, sit, table, blue, speech	mall, tie, interview, sunglasses, give
	1	news, wear, tie, business suit, man	business suit, tie, suit, interview, microphone	bookshelf, office, interview, black, tie
	2	stand, mask, man , wear, person	mask, equipment, syringe, hand, liquid	worker, tablet, syringe, hospital room, garment
Ukraine war	0	business suit, speech, stand, tie, man	speech, podium, give, deliver, microphone	give, podium, business suit, speech, tie
	1	business suit, wear, tie, stand, man	business suit, news, interview, woman, suit	shake, suit, interview, business suit, tie
	2	man, rubble, debris, city, building	debris, city, rubble, destruction, damage	explosion, flame, damage, garbage, rubble
	3	stand, army, camouflage, person, man	camouflage, army, soldier, gun, equipment	weapon, load, gun, soldier, rifle

Table 1: Results from concept tagging

3.2 Identifying Visual Concepts

Prior to any comparative analysis, in order to control for the image content, the thumbnails are first divided into subgroups based on concepts (i.e. visual themes). Two methods are explored for this procedure. First, we use PlacesCNN [31], a neural network pre-trained for scene recognition to convert images into location embeddings, a 365-dimensional probability vector corresponding to different scene locations. Using these embeddings, clustering methods (K-Means clustering and hierarchical agglomerative clustering) are applied to group together thumbnails that are similar in content. The second approach involves Concept [32], a CLIP and BERTopic-based open source tool that performs topic modeling on images by first encoding images using CLIP and then identifying clusters on top of those embeddings. An example of the results produced using this method is shown in 3. For this experiment, the second approach is used, the main reason for this choice being that the location classes from the PlacesCNN model are very specific and this can lead to thumbnails shot in similar settings not being clustered together. From qualitative inspection, the clusters from Concept appear less affected by minor differences in image content, which is more favorable given the relatively small size of the dataset being investigated.

$$w_{i,j} = tf_{i,j} \log\left(\frac{N}{df_i}\right) \tag{1}$$

Thumbnail clusters themselves carry no easily interpretable meaning, so to understand the visual concepts being extracted we associated each cluster with 5 textual tags. To do this, we leverage Tag2Text [33], a model capable of recognizing 3,429 commonly human-used categories, to generate tags for every image. Then, on a cluster level, we aggregate these tags to generate 5 tags that are most representative of the cluster's concept. For this, three approaches are experimented with. First, use the top 5 most commonly occurring tags. Second, use the same approach as the first but with the added requirement that these tags appear less than some k times across all images (including other clusters). Third, computing a tf-idf score for each tag with respect to a cluster using equation 1, and use the 5 tags with the largest weights. Under our setup, $w_{i,j}$ is the tf-idf weight of tag i in cluster j, $tf_{i,j}$ is the frequency of the tag in the cluster, N is the total number of clusters, and df_i is the number of clusters containing the tag. The second and third approaches were introduced to filter out certain tags, such as `man`, that appear in almost every thumbnail and thus are not discriminative against clusters. The results from these tagging methods respectively are summarized in table 1, and the tags generated using approach 2 are used.

3.3 Extracting Image Features

Image features are extracted using the hand-crafted approach, where various pre-defined statistical measures are computed for each image. An exhaustive list of features, including references to their use in the literature is displayed in table 2. A majority of our feature set is adopted from Redies et al. [17], though *edge orientation entropy* is excluded for practical reasons. Additionally, we include two additional features *shot scale* and *bag-of-objects*. A brief explanation of each feature is included here, but the reader is encouraged to consult the references for more details.

Color Features. *Hue, Saturation, Lab(a), Lab(b), Color Entropy.* Color is an important part of images especially when it comes to aesthetics. The use of colors can affect the mood of an image, certain colors carry certain connotations, and more, making it a commonly assessed aspect in image-based tasks. There are multiple ways to represent colors numerically, RGB, where the color of each pixel is described using three numbers corresponding to red, green, and blue, being the most popular. However, alternative color spaces exist, such as HSV (hue, saturation, value) and Lab (lightness, a, b), both of which have been shown to be more aligned with the human perception of color [34]. *Hue* and *saturation* are measured as the mean value across all pixels of the H, and S channels of the image represented in HSV space. Similarly, *lab(a)* and *lab(b)* are the mean values of the A, and B channels respectively in LAB space. Finally, *color-entropy* is introduced to capture the 'colorfulness' of an image, and is measured as the Shannon entropy of the hue channel for a given image. Images with many different values of hue across its pixels will have high entropy whereas ones with predominantly one or few colors will have low entropy.

Image Dimension Features. *Aspect Ratio, Image Size.* Although not as important in this study, considering thumbnails are reshaped into a pre-defined dimension by YouTube, for completeness we measure the *aspect ratio* and *image size* of the thumbnail. *Aspect ratio* is defined as the ratio between the height and width of the image, calculated by dividing the image width by the image height, capturing how horizontally dominated the image dimensions are. *Image size* is calculated as the sum of image width and image height. It is necessary to include both as images can have the same aspect ratio while differing in image size, where one is a proportional scale up/down of the other. Image width and image height are measured in pixel units. While thumbnails are a pre-defined dimension, it is important to note that both of these features can still vary within our dataset for two reasons. First, YouTube videos can be one of two content types, videos or shorts, which are presented slightly differently. Second, since we automatically trim any black bars as part of the preprocessing, thumbnails with black bars that are added by the user, in addition to YouTube's automatic padding, will cause slight variation.

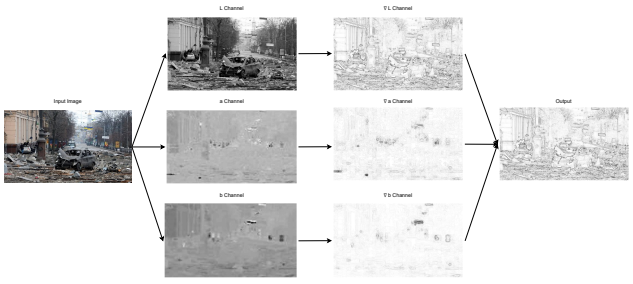

Figure 4: Process of calculating the gradient

Light Features. *Contrast, Luminance, Luminance Entropy.* To complement color, several features for lightness (i.e. brightness) are also considered since color alone doesn't capture everything. For instance,

Group	Literature Reference(s)	Dimensions	Feature	Description
Color	Li et al. 2006 [35], Li & Chen 2009 [36], Mallon et al. 2014 [37], Thieleking et al. 2020 [38], Schifanella et al. 2015 [39], Iigaya et al. 2021 [40], Geller et al. 2022 [41]	5	*Hue*	mean of hue channel in HSV space
			Saturation	mean of saturation channel in HSV space
			Lab(a)	mean of a channel in LAB space
			Lab(b)	mean of b channel in LAB space
			Color Entropy	Shannon entropy of hue channel in HSV space
Image Dimension	Li et al 2006 [35], Mallon et al. 2014 [37], Iigaya et al. 2021 [40]	2	*Aspect Ratio*	image width divided by image height (pixels)
			Image Size	image width + image height (pixels)
Light	Peli 1990 [42], Schifanella et al. 2015 [39], Sidhu et al. 2018 [43], Mater 2018 [44], Iigaya et al. 2021 [40]	3	*Contrast*	standard deviation of L channel in LAB space
			Luminance	mean of L channel in LAB space
			Luminance Entropy	Shannon entropy of L channel in LAB space
HOG	Bosch et al. 2007 [45], Braun et al. 2013 [46]	3	*Self-Similarity*	similarity of HOG features
			Complexity	mean gradient strength
			Anisotropy	standard deviation of HOG features
Fourier	Redies et al. 2008 [47]	2	*Fourier Slope*	slope of line of best fit on the log-log plot of the Fourier power spectrum
			Fourier Sigma	RMSE of line of best fit on the log-log plot of the Fourier power spectrum
Symmetry	Brachmann & Redies 2016 [48], Bertamini & Rampone 2022 [49]	2	*Symmetry-lr*	left-right symmetry based on first layer activations on pre-trained AlexNet
			Symmetry-ud	up-down symmetry based on first layer activations on pre-trained AlexNet
CNN	Brachmann et al. 2017 [50]	2	*Sparseness*	median variance of each max-pooled response map from the first layer of pre-trained AlexNet
			Variability	variance over all max-pooled response maps from the first layer of pre-trained AlexNet
Other	Savardi et al. 2018 [51], Zhang & Peng 2022 [26]	>1	*Shot Scale*	close, medium, or wide shot based on fine-tuned CNN
			Bag of Objects	frequency of objects detected by YoloV5

Table 2: Summary of visual features extracted

the effect of viewing a very bright vibrant red, such as Salmon, may not be equivalent to that of a darker red such as Burgundy. *Luminance*, which describes how bright an image is overall, is measured as the mean value of the L channel in LAB space. In this paper, *contrast*, a concept that often arises in the context of Photography, is calculated as the standard deviation of the L channel. Images with many varying light and dark spots will have high contrast whereas images where the brightness is relatively unchanged throughout will have low contrast. *Luminance entropy* is the Shannon entropy of the L channel and describes roughly the same concept as contrast but with some small differences.

$$g = \sqrt{g_x^2 + g_y^2} \tag{2}$$

$$\theta = \arctan \frac{g_x}{g_y} \tag{3}$$

Histogram of Oriented Gradients Features. *Self-Similarity, Complexity, Anisotropy.* A histogram of orientated gradients (HOG) is a feature descriptor that was originally introduced for object recognition [45]. The method describes the orientation of gradients within localized regions of an image, storing them as a probability density histogram where the bins are defined by the angle and the weight is determined by the magnitude of the gradient. To calculate the gradient, x and y Sobel filters, which capture the first derivatives, are applied to each channel (L, a, b) of the image, producing g_x, g_y respectively. Then, the total gradient g and angle θ can be determined using the equations above. Finally, the three channels are merged together by taking the maximum gradient and corresponding angle for each pixel. The process up until this point is depicted in figure 4.

Self-similarity measures the similarity of HOG features in a smaller, localized patch of an image with larger regions. There are several ways to calculate self-similarity which differ in the granularity at which these patches are considered and how the larger patch that each smaller patch is compared to is selected. In this study, we calculate self-similarity using neighboring patches on a level of 3 (other variations are described in full in Braun et al. [46]). First, an image is divided evenly into four quadrants, which are each divided again into even quadrants, and so on. In our case, this is repeated 3 times, i.e. the level we are performing self-similarity at. The result is an even 8x8 partition of the image with 64 patches in total. Then, for each patch, a similarity score between its HOG feature (represented as a d-dimensional vector where d is the number of bins) and each adjacent patch is calculated using the Histogram Intersection Kernel outlined in equation 4, where $h, h^{'}$ are two HOG vectors, $h_i, h_i^{'}$ are the corresponding i-th entries, and n is the number of bins used in the histogram. The overall *self-similarity* metric is the median similarity score between all patch, neighboring patch pairs in the image. *Complexity* is calculated as the mean gradient strength throughout an image. An image with lots of sharp edges and therefore very strong gradients would have high complexity and vice versa. Finally, *anisotropy* measures the relative magnitude

of gradients with respect to different orientations, calculated as the standard deviation of HOG features. Low values of *anisotropy*, and thus lower standard deviations, indicate a more uniform strength across all orientations (bins) whereas a high value would suggest strongly differing strengths at different orientations. It describes whether there is a balance of edges with respect to the angle of orientation or not throughout an image.

$$\mathrm{HIK}\left(h, h^{'}\right) = \sum_{i=1}^{n} \min(h_i, h_i^{'}) \tag{4}$$

Fourier Transformation Features. *Fourier Slope, Fourier Sigma.* The Fourier transformation is a commonly used technique in image processing that maps an image to the frequency domain, decomposing the image into a sum of periodic sine and cosine components. An example of the transformed image is supplied in figure 5, where the coordinates represent the frequency of the components (lower frequency components are closer to the center) and the color represents the magnitude of that component to the image. Low-frequency components are larger, less variant parts of the image whereas high-frequency components capture sharp details or rapid changes in intensity. From here, we calculate the power (magnitude squared) of the Fourier transformation as a function of the radial average from the center taken at 1-pixel intervals, which describes the relative decrease in presence from low-frequency components to high-frequency components. The log plot is then computed and *Fourier slope* is calculated as the slope of the line of best fit whereas the *Fourier sigma* is calculated as the root mean squared error of the line of best fit. The former describes how steep the drop-off is, ie. the relative strength of low-frequency components compared to high-frequency components, whereas the *Fourier sigma* describes how much this decrease deviates from a linear course. An image with larger *Fourier slope* (less negative) suggests a stronger presence of details whereas smaller values suggest an overall smoother image.

Symmetry Features. *Symmetry-lr, Symmetry-ud.* Pre-trained models, such as AlexNet, that have been trained on large datasets often learn certain low-level features such as color and texture in their initial layers before advancing to more abstract, complicated features in later layers. Previous work has demonstrated that using activation maps from these initial layers can produce a symmetry metric that is more closely aligned with human perception

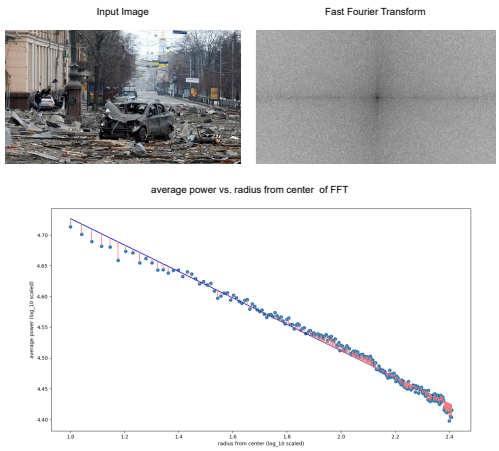

Figure 5: Process of calculating the Fourier features

[48]. To calculate symmetry, these first layer AlexNet activations for an image I_l and a flipped copy F_l are used as shown in equation 5. The term on the right sums over all first layer filters f and all pixel coordinates x, y and calculates the difference between those from the original vs. flipped activations. This number describes asymmetry, and so the final score is calculated by taking $1-$ this value. A horizontal flip is used to calculate the *lr* (left-right) symmetry whereas a vertical flip is used for *ud* (up-down) symmetry.

$$S(I_l, F_l) = 1 - \frac{\sum_{x,y,f} |I_l(x,y,f) - F_l(x,y,f)|}{\sum_{x,y,f} \max\left(I_l(x,y,f), F_l(x,y,f)\right)} \tag{5}$$

CNN Features. *Sparseness, Variability.* First layer AlexNet activations are also used to calculate *sparseness* and *variability.* *Sparseness* is defined as the median of the variances for each resulting response map (max-pooled). Images with low sparseness indicate less variance in response maps, meaning many filters respond to a similar degree in different max-pooled patches of the image, and thus the image is richer. Conversely, images with high sparseness often correspond to images with homogeneous patches, hence the term sparse. *Variability* is measured as the variance over all response maps, and captures the inverse of *self-similarity.* If an image has high variability, then low-level features across different subregions are very diverse, whereas an image with low variability suggests the presence of very similar features across subregions.

Other Features. *Shot Scale, Bag-of-objects.* To complement aesthetic features, we introduce two ad-

ditional features *shot scale* and *bag-of-objects*. From CineScale, a project aimed to recognize cinematic features with machine learning, we use a trained CNN to predict the shot scale of an image as close, medium, or long [51]. In addition, using YoloV5, a lightweight open-source model with object detection capabilities, we represent each image using an object embedding, where each dimension represents the presence of a certain object and the value corresponds to the frequency of that object in the image (analogous to bag of words model in NLP). The dimensionality of this vector is dependent on the dataset and varies between news stories. We set the threshold confidence for object detection at 0.75.

3.4 Sanity Checks

We first perform a few sanity checks (section 4.1) before proceeding with the analysis of our research questions (section 4.2 - 4.5). First, we evaluate the correctness of our concept modeler by considering to what extent the contents of the images within a cluster are consistent with one another. Using the previously mentioned PlacesCNN model, which can label each image generally as indoor/outdoor, we measure the entropy of each concept cluster using the Gini index as defined in equation 6, where c can be one of two classes (indoor/outdoor) and p_j is the proportion of images in the concept that belongs to class j. We expect the entropy to be low (relative to the entropy before concept modeling), with images within the same cluster sharing similar general settings. Second, we qualitatively evaluate the extraction of CNN features by visualizing images with the highest/lowest values for certain features. Third, we investigate whether *variability* is inversely related with *self-similarity* as claimed in the literature by applying Spearman's rank correlation test to quantify the direction of the relationship between these two features and its strength.

$$I_g = 1 - \sum_{j=1}^{c} p_j^2 \tag{6}$$

3.5 Cross-cultural Analysis

To evaluate cross-cultural differences between Western vs. Chinese thumbnails, we formulate four key research questions,

RQ1. How do Western and Chinese viewership behaviors differ on YouTube?

The distribution of viewership statistics (views, likes, comments) is compared between US and Chinese news channels. Video statistics are collected, normalized by channel, and then converted to a histogram based on frequency. Following the power law, a log-log plot of these histograms is computed for each channel, and a line of best fit is fitted on each. The slopes of these lines are compared with each other to highlight differences in how quickly the frequency of videos drops off with increasing views, likes, and comments.

RQ2. Do Western and Chinese channels tend to favor covering different concepts (i.e. news subtopics)?

For each news channel, the frequency of thumbnails per visual concept is computed. All sources contain an even number of thumbnails per news story, so no additional normalization is needed. To see whether news channels differ in their distributions, these frequency distributions are plotted against each other on a bar plot. Additionally, we group thumbnails into bins based on the date the video was published, and then see how thumbnail concepts vary over time for each channel.

RQ3. What are the notable differences, if any, between Western vs. Chinese thumbnail aesthetics?

First, a feature correlation matrix between each image feature is computed across the entire dataset. Then, for each news event, and concept, the thumbnails are split by region (Western/Chinese) according to channel. For each aesthetic feature, a two-sample t-test is applied to test the significance of the difference between Western and Chinese thumbnails. These results, as well as the average percent difference in the average feature value (indicating which culture tends to exhibit greater values for a feature and to what degree), are aggregated and displayed in a summary table. The use of different shot scales is compared using a bar chart for each event, and concept and the results from the bag-of-objects embedding are compared by word clouds.

RQ4. What is the relationship between thumbnail aesthetics and video performance metrics, and does this change between Western and Chinese audiences?

Similar to the previous section, for each story, concept, and culture, we evaluate the aesthetic features of a thumbnail along with the video's views, likes, and comments. Views are min-max normalized on the channel level and likes and comments are normalized by the views on the video (i.e. like/comment rate).

Then, Spearman's rank correlation is used to obtain a p-value indicating whether an aesthetic feature has a significant influence on a performance metric or not. These are compared across cultures and notable similarities/differences are reported.

4 Results

4.1 Sanity Checks

Concept modeling clusters based on some notion of image content. After processing and cleaning the dataset, the Ukraine news story, had 599 outdoor images and 575 indoor images (Gini index of 0.50) whereas the COVID-19 news story had 239 indoor and 931 outdoor images (Gini index of 0.33). The disparity is likely due to the latter containing a lot of images at press conferences and indoor medical settings whereas the former was more diverse with shots of rubble, outdoor military operations, in addition to conference rooms. Note that there are fewer than the expected 2400 images in total as a result of errors when extracting features, however, the distribution across channels remained relatively even. Using the indoor/outdoor classification of PlacesCNN as proxies of image content, the Gini indices were obtained. For COVID-19, the indices of the 3 concepts were 0.23, 0.38, and 0.16. Two clusters resulted in more 'pure' image sets, whereas one, that corresponded to a medical theme ("mask", "equipment", "syringe" tags), was slightly more mixed than before. This is likely due to a mix of images inside hospitals but also those of people getting tested outdoors or in drive-thru tests. Since Ukraine's Gini index was already at maximum entropy, clusters could only have indices equal to or less than 0.5. In one cluster it remained around maximal entropy, in two it decreased slightly, and in the last one, the cluster relating to "man", "rubble", and "debris" as themes, the Gini index dropped significantly to 0.26. This is likely due to the fact that these shots are almost always outdoors. Our clustering method seems somewhat consistent with the general predicted setting of the scene.

CNN-based features capture what they intend to. Thumbnail examples for images that ranked at the top/bottom for three CNN-based features, *symmetry-lr*, *sparseness*, and *shot scale*, are given in figure 7. Images with low left-right symmetry tend to have a dominant subject (person, or data figure) framed at the edges of the screen with empty space on the other side whereas those with high symmetry do indeed portray an evenness on both ends of the frame. Sparse images tend to be dull, faded-out thumbnails whereas those less sparse are more complex in subjects, edges, etc. Images with high shot scale (i.e. wide shots) are scenic images capturing large lands whereas those identified as close-up shots are often portraits of people. Through qualitative inspection, our features seem to be accurate reflections of the visual aspects they attempt to describe.

Variability is inversely related to self-similarity and vice versa. We fit a Spearman's correlation test on the two features across all images in our dataset to obtain a correlation coefficient of -0.14 with $p = 3.1 * 10^{-12}$. Though the magnitude of the correlation coefficient is not relevant, since the two features are not necessarily on comparable magnitudes, the negative sign indicates an inversely related relationship, validating our assumption. Furthermore, the p-value is very small. By convention, we benchmark significance at $p = 0.05$, confirming that this inverse relationship is significant within our dataset.

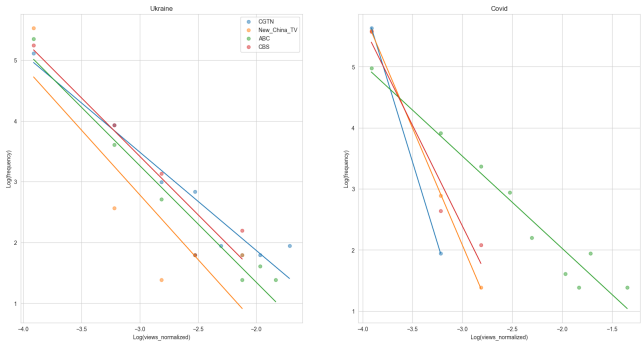

Figure 6: Log-log plot of views by channel

4.2 Viewership Behaviors

The log-log plot of the frequency of videos and the views, min-maxed normalized by channel, separated by news story are shown in figure 6. Data points where the frequency was 0 were omitted to smooth the data since the log of 0 is undefined. The behavior of views by video seems to adhere to the power law, with many videos having very few views and falling off quickly as the number of views increases, resulting in a linear pattern on the log-log plot. Though all channels follow roughly the same behavior for the Ukraine event, in the COVID-19 news story, the slope of the line of best fit for ABC is much greater than the other channels, suggesting that ABC COVID-19 videos have a slightly more even distribution of views, compared to other channels that have a larger head. This may be due to less frequent uploads or a more dedicated, regularly returning audience.

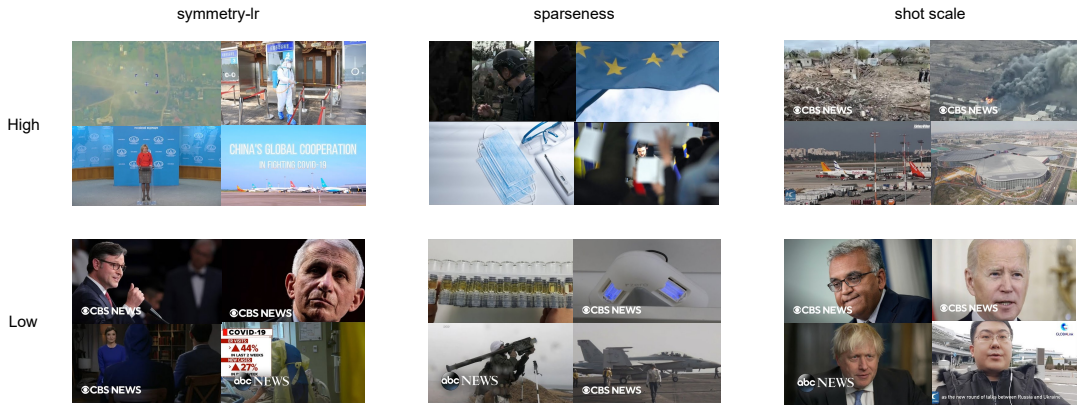

Figure 7: Visualization of top and bottom ranking images for CNN-based features

Likes don't adhere to the power law distribution as much as views. Thus, we choose to display the results (figure 8) in histogram form, where videos are binned by like rate, the number of likes on a video divided by its views, and the number of videos per bin is calculated. There is an observable difference between audiences, namely while Western channels have a large head, with a majority of the videos having a low like rate, Chinese channels tend to peak slightly later on, around 0.07, diverging from the power law. Chinese audiences are more likely, on a view-by-view basis, to like the video. This may indicate that Chinese audiences may be more active participants in videos, directly interacting with them by clicking the like button. Alternatively, it is also important to note that Western channels have significantly more views overall, meaning they are more passive audiences which also may drown out the like rate more (despite likes already being normalized by views). There are no detectable differences in the rate at which viewers comment on a video.

4.3 Preferences for Visual Concepts

The Western and Chinese distribution of videos across visual concepts is displayed in figure 9. For the Ukraine news story, Western thumbnails have a noticeably stronger focus on militaristic visual themes, with more YouTube thumbnails falling under the "army, camouflage" and "rubble, debris" concepts compared to Chinese channels. The former concept is mainly populated by images of soldiers wearing camouflage uniforms outdoors sometimes alongside military vehicles and the latter contains images of buildings in the active war zone that have been damaged by explosives. The other two concepts for Ukraine, are focused on the geopolitics of the conflict, containing images of political figures in conference rooms, government officials giving speeches at podiums, or headshots of news reporters reporting on the conflict. It is evident that Western sources tend to cover the militaristic aspect of the conflict, reporting photographs of the battle, whereas Chinese sources have a stronger preference for the geopolitical nature. Both cultures have a large emphasis on the medical context of COVID-19, with the "mask, equipment, syringe" concept being by far the most common theme. The distribution for COVID-19 concepts is roughly similar across cultures.

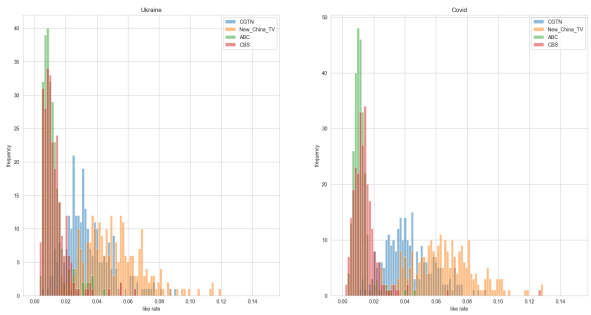

Figure 8: Like rate distribution across events, channels

This difference may reflect the intentions of the different media outlets or the different audience preferences. Chinese channels, due to political affiliation with Russia, who are commonly seen as the aggressor of the conflict, might be more hesitant to cover the militaristic point of view, especially from the Ukrainian point of view. On the other hand, sources based in the US, which have openly declared their support for Ukraine, may be more focused on bringing awareness to the damage the war is causing (hence the "rubble, debris" images) or building up a narrative to support the Ukrainian side depicting them as bravely defending their nation (hence the "soldier, camouflage" images). Alternatively, another hypothesis is that they reflect the audience's preferences. Western

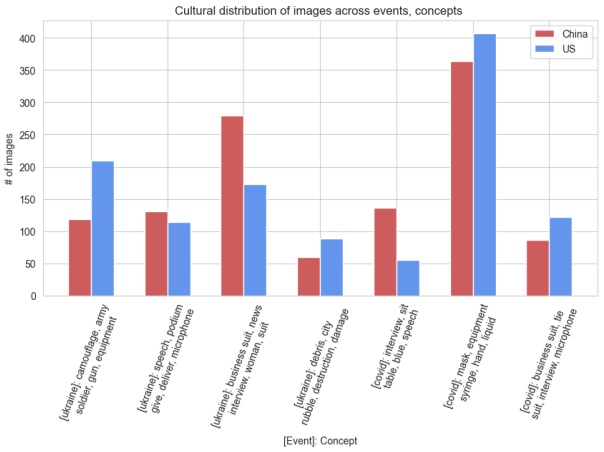

Figure 9: Distribution of videos across concepts grouped by culture

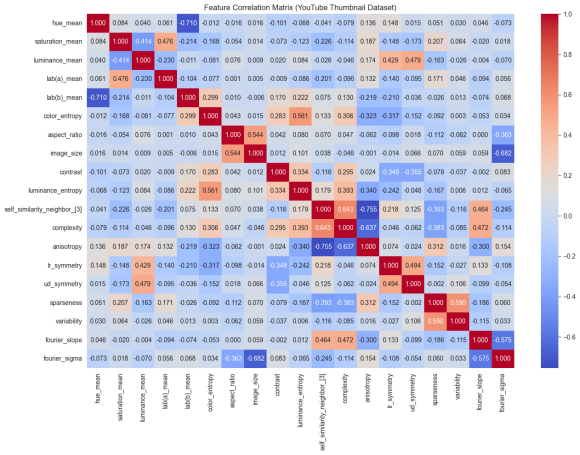

Figure 10: Correlation matrix for aesthetic features

audiences, out of curiosity, may be more drawn to click on thumbnails that involve violence and so channels might cater to this as a means of increasing their metrics. For Chinese channels these metrics might not be as important since in China, YouTube is not the main video platform, and likely not a large source of revenue. In fact, very few Chinese-based news YouTube channels exist, and therefore they might not be strong competition for views on the platform.

A more comprehensive view, showing the presence of thumbnails per concept over time is shown in figure 11. For both events, there appear to be spikes where certain months many videos tend to be uploaded. In March 2022, there was a sudden increase in video uploads related to COVID-19, coinciding with a period when a new COVID-19 variant, the omicron variant, began spreading rapidly. Another example is for the Ukraine event, there is a small influx of videos in the second month of 2022, likely corresponding to the United Nations General Assembly that took place to address the situation in Ukraine. For Chinese YouTube channels, of the 300 videos scraped, most seem to be uploaded around these peaks, with very little activity outside of these time periods. US channels, however, tend to have a slightly more consistent upload schedule, regularly uploading videos and so they are not as concentrated in certain points. This may be related to the commercial interests of US channels, trying to drive viewer engagement regularly whereas Chinese YouTube channels may be less business-oriented.

4.4 Cultural Differences in Thumbnail Aesthetics

Assessing the feature correlation matrix, there are a couple of features that exhibit correlation with each other. The mean value of the b channel in LAB-space is inversely correlated with the mean hue (h channel in HSV-space). This may be attributable to how increasing the value of the b channel results in a transition from blue colors to green-yellow-red colors (depending on the other a dimension) and that green and yellow colors are defined to be lower in h value than blue colors in HSV space. Another strongly correlated pair of features is self-similarity and anisotropy, likely because if an image has high anisotropy, with a greater standard deviation of edge intensity across orientations, the image is likely more complex in the direction of edges and thus the histogram of oriented gradients from image patches are more likely to be dissimilar adjacent patches rather than consistent. An interesting correlation is the inverse relationship between image size and Fourier sigma, which might be a by-product of the fact that some Thumbnails, due to varying sizes of black bars, have slightly different dimensions and that might be affecting how the metric is calculated. Other histogram of oriented gradient features tend to show some correlation with each other as well. The full correlation matrix is shown in figure 10.

To examine cultural preferences for image aesthetics in YouTube thumbnails, we display the results from a two-sample (Chinese, Western groups) t-test for each feature on each event, and concept subgroup. The value in each cell reflects the percentage difference between the feature averages for each culture, where a negative value indicates Chinese thumbnails exhibit a larger value for the feature on average and vice versa. The cell colors reflect these values, with red indicating Chinese thumbnails tend to have larger

Figure 11: Distribution of thumbnail concepts over time

feature values and blue indicating US thumbnails tend to have larger feature values. Bolded values represent features, events, and concepts in which the difference between Chinese and Western thumbnails is significant ($p < 0.05$). Many features seem to generalize across events, and concepts, that is one culture tends to have larger values than the other irrespective of the image content. Examples of Chinese and US thumbnails that are representative of certain features are provided throughout this section.

Color features that show a discrepancy between culture groups include hue, saturation, and luminance. Chinese thumbnails tend to have greater hue values than Western thumbnails, which, according to the results, is especially true in the context of thumbnails depicting interviews/press conferences. Qualitatively, thumbnails with high hue values tend to correspond to images with a greater presence of blue, whereas those with low hues tend to be images with warmer colors such as yellow, orange, and green. In the case of thumbnails of interviews from COVID-19, blue hues often arise from images taken of government officials giving press conferences, where the choice of backdrop is very often a blank, strong blue (figure 13). Though backdrops are present in Western thumbnails as well, they are usually more variable, use a less vibrant shade of blue, and often contain patterns such as logos. There are also generally fewer Western thumbnails taken from interviews/press conferences with set backdrops and more where the background is often a natural scene such as a room or the white

event	concept	hue_mean	saturation_ mean	luminance_ mean	lab(a)_mean	lab(b)_mean	color_entropy	aspect_ratio	image_size	contrast	luminance_ entropy	self_similarity_ neighbor	complexity	anisotropy	lr_symmetry	ud_symmetry	sparseness	variability	fourier_slope	fourier_sigma
ukraine	camouflage, army, soldier, gun, equipment	-0.19	0.14	-0.15	-0.01	0.02	0.02	0.01	0.00	0.05	0.00	-0.01	0.05	-0.07	-0.18	-0.14	0.12	-0.24	-0.01	0.00
ukraine	speech, podium, give, deliver, microphone	-0.25	-0.04	-0.05	0.01	0.04	-0.02	-0.02	0.01	0.12	0.01	-0.04	-0.05	0.05	-0.15	-0.10	0.24	-0.28	0.08	0.01
ukraine	business suit, news, interview, woman, suit	**-0.12**	**0.37**	**-0.20**	**0.02**	**0.02**	-0.01	0.01	0.00	0.00	0.00	**-0.08**	**-0.15**	**0.06**	**-0.29**	**-0.13**	**0.34**	-0.18	**0.06**	**0.01**
ukraine	debris, city, rubble, destruction, damage	-0.03	0.13	-0.01	0.01	0.00	-0.01	0.02	0.01	0.02	0.00	**-0.04**	-0.08	0.02	**-0.18**	**-0.09**	0.10	-0.28	**0.12**	**0.01**
covid	interview, sit, table, blue, speech	**-0.16**	-0.08	**-0.23**	0.00	**0.08**	**0.08**	0.00	0.00	**0.12**	0.01	0.02	0.05	-0.03	**-0.30**	**-0.16**	**0.21**	-0.04	**0.07**	**0.01**
covid	mask, equipment, syringe, hand, liquid	**-0.11**	**0.09**	**-0.07**	**0.02**	**0.01**	-0.01	**0.01**	**0.01**	**0.07**	0.00	**-0.07**	**-0.11**	**0.08**	**-0.19**	**-0.10**	**0.25**	**-0.14**	**0.11**	**0.01**
covid	business suit, tie, suit, interview, microphone	0.04	**0.18**	**-0.19**	0.00	0.01	0.01	-0.01	0.00	0.03	**0.03**	-0.02	0.07	-0.05	**-0.23**	**-0.22**	**0.24**	-0.05	-0.02	0.00

Figure 12: Cultural comparison of image aesthetics (red Chinese, blue Western)

house. Chinese images of press conferences about Ukraine also tend to use a blue backdrop, which, based on the text, seems to be the default choice for press conferences for the "Ministry of Foreign Affairs". In the medical context of COVID-19, both cultures display the blue tones that commonly appear in hospital garments, though the results suggest Chinese thumbnails emphasize the blue colors from these scenes slightly more. In addition to blue hues, large hue values can also arise from 'cool' tones as opposed to 'warm' tones on an image. Chinese thumbnails of regular, outdoor, scenic images of Ukraine as a city (unrelated to the military or rubble, debris concept), largely depict a blue sky, giving these images a 'cool' tone.

Figure 13: Chinese press conference thumbnails

Alongside color, Chinese thumbnails tend to also be brighter on average. One of the reasons for this is that Chinese thumbnails seem less formal overall, occasionally even overlaying animated components on top of images to create thumbnails for special re-occurring series (figure 16). This use of animated features tends to result in a bright and vibrant image. This is rarely the case for US channels which almost exclusively use real photographs as thumbnails (sometimes a data-centric graphic will be employed). Additionally, Western thumbnails, which are often portrait shots, appear edited in a cinematic way, such that the subject's face is clear and sharp, but the background is often darkened which can cause them to be darker on average. The use of portrait shots and the darkening of the background might suggest that Western channels place more importance on the specific individual compared to Chinese channels. Also, Western images, though they seem to be less 'colorful' according to hue averages tend to be more saturated, using stronger, deeper colors. This is possibly a result of the editing process whereas Chinese thumbnails seem more candid as if they were taken directly from a frame in the video rather than photographed and edited separately.

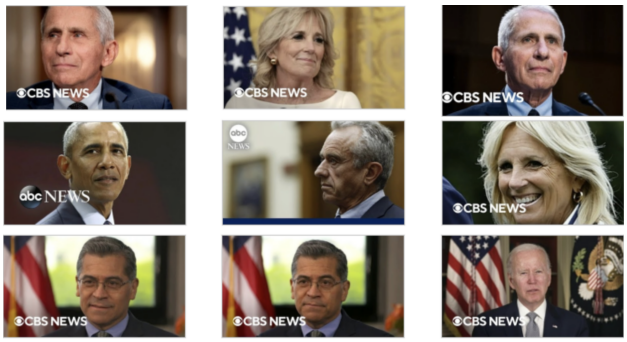

Figure 14: US press conference thumbnails

In addition, Chinese thumbnails are, on average, more symmetric in both the horizontal and vertical directions. According to both symmetry metrics, Chinese images were more symmetric across all events, and concepts, with this difference being significant in 5 out of 7 subtopics. One potential explanation is that Western thumbnails adhere more closely to the photographic principle of the rule of thirds where, for instance, in portraits, the subject figure is framed to the left or the right of the frame with negative space taking up the remaining two-thirds of the image, which may create asymmetry. Chinese thumbnails, on the other hand, frame the subject directly in the middle of the frame, which may be a result of the fact that Chinese thumbnails seem to come from the actual video footage itself, rather than

culture	metric	event	concept	hue_mean	saturation_mean	luminance_mean	lab[b]_mean	lab[b]_mean	color_entropy	aspect_ratio	image_size	contrast	luminance_entropy	similarity_neighbo	complexity	anisotropy	lr_symmetry	ud_symmetry	sparseness	variability	fourier_slope	fourier_sigma
China	views	ukraine	camouflage, army,	0.16	0.03	0.00	0.00	-0.09	-0.04	0.09	-0.11	-0.08	-0.16	0.00	-0.14	0.09	0.16	0.12	-0.06	-0.06	-0.18	0.15
		ukraine	speech, podium, giv	0.03	0.07	-0.05	0.09	-0.08	-0.12	-0.07	0.04	-0.03	-0.04	-0.03	-0.15	0.11	0.01	0.00	0.14	0.02	-0.07	0.09
		ukraine	business suit, news,	0.06	0.01	0.02	0.08	-0.03	0.12	0.09	-0.06	**0.13**	**0.14**	-0.07	-0.04	0.01	-0.04	0.00	-0.01	-0.03	**-0.16**	**0.16**
		ukraine	debris, city, rubble,	-0.04	0.12	0.03	0.19	0.09	-0.05	**-0.30**	**0.32**	0.00	0.15	0.22	0.23	-0.20	-0.10	0.25	-0.16	-0.02	-0.08	0.25
		covid	interview, sit, table	-0.09	0.00	**-0.17**	**0.28**	0.15	**0.20**	0.00	-0.05	**0.17**	0.09	0.02	0.11	-0.12	-0.17	-0.15	-0.01	-0.01	-0.11	**0.17**
		covid	mask, equipment, s	0.04	-0.10	**0.12**	-0.03	-0.05	**-0.16**	-0.02	-0.08	-0.09	**-0.20**	-0.08	**-0.15**	**0.13**	**0.13**	0.00	-0.03	-0.04	-0.03	0.01
		covid	business suit, tie, s	-0.07	-0.09	0.15	0.12	0.00	-0.05	-0.10	0.10	-0.18	-0.02	0.08	0.04	-0.03	0.12	**0.32**	-0.15	-0.01	-0.18	**0.23**
	likes	ukraine	camouflage, army,	0.02	-0.17	0.16	-0.05	-0.08	0.03	-0.13	0.13	0.12	0.21	-0.02	0.11	0.02	-0.07	0.03	0.05	0.06	0.13	-0.09
		ukraine	speech, podium, giv	0.04	0.08	-0.32	0.00	-0.06	0.14	0.00	0.04	-0.09	-0.04	-0.06	0.04	-0.20	-0.16	-0.13	0.21	0.22	-0.02	0.00
		ukraine	business suit, news,	0.08	-0.05	-0.03	-0.08	-0.05	**-0.13**	0.05	-0.07	**-0.14**	**-0.15**	0.08	-0.02	-0.05	0.09	0.04	0.01	0.02	0.12	**-0.16**
		ukraine	debris, city, rubble,	0.15	-0.18	0.00	**-0.26**	-0.14	0.17	0.11	-0.18	0.03	-0.19	-0.12	-0.20	0.09	0.06	-0.17	0.08	0.00	0.08	-0.25
		covid	interview, sit, table	-0.14	-0.07	**-0.21**	0.16	**0.23**	0.10	0.16	**-0.22**	0.16	-0.09	-0.03	-0.04	0.00	-0.16	-0.03	0.00	-0.10	-0.09	0.03
		covid	mask, equipment, s	0.05	0.10	0.01	-0.02	**-0.13**	-0.05	-0.01	-0.04	0.05	-0.02	**-0.14**	-0.08	**0.16**	0.03	-0.05	0.05	0.01	-0.07	0.06
		covid	business suit, tie, s	0.19	-0.19	0.14	0.01	-0.12	0.19	0.05	-0.05	0.01	-0.02	0.00	-0.12	-0.07	0.14	-0.15	0.05	-0.09	0.02	-0.05
	comments	ukraine	camouflage, army,	0.07	-0.04	0.10	-0.03	-0.11	0.11	0.08	-0.04	0.21	0.04	-0.03	0.03	0.05	0.05	-0.03	-0.09	-0.07	-0.05	0.04
		ukraine	speech, podium, giv	0.06	0.33	-0.21	0.26	-0.17	-0.01	-0.09	0.13	-0.14	-0.21	-0.17	-0.09	0.00	-0.09	-0.11	0.27	0.26	-0.12	0.10
		ukraine	business suit, news,	0.06	0.03	-0.01	0.02	-0.06	-0.09	-0.05	0.02	**-0.13**	**-0.13**	-0.07	-0.11	0.09	0.04	**0.13**	0.04	-0.03	-0.09	0.05
		ukraine	debris, city, rubble,	0.17	-0.18	-0.05	**-0.34**	**-0.25**	0.04	0.12	0.03	0.13	0.08	0.17	0.11	**-0.21**	0.03	0.08	-0.10	-0.03	0.21	**-0.27**
		covid	interview, sit, table	-0.01	0.02	-0.07	0.09	-0.05	-0.12	0.00	-0.06	0.07	-0.06	-0.01	0.05	-0.06	-0.06	-0.10	0.14	0.14	-0.11	0.12
		covid	mask, equipment, s	0.04	0.08	-0.04	-0.02	-0.07	-0.07	-0.04	0.09	0.04	**0.14**	0.01	0.07	-0.06	-0.06	-0.05	0.04	0.07	0.02	-0.02
		covid	business suit, tie, s	0.00	-0.04	0.17	-0.11	-0.06	0.07	0.07	-0.07	0.04	-0.09	**-0.24**	-0.09	0.17	0.09	0.00	-0.05	-0.09	-0.01	0.00
US	views	ukraine	camouflage, army,	-0.09	-0.37	0.07	-0.17	-0.06	-0.11	0.30	-0.28	-0.05	-0.03	0.15	0.02	-0.06	0.09	-0.03	-0.09	-0.07	-0.03	0.05
		ukraine	speech, podium, giv	0.15	-0.08	-0.06	-0.08	-0.18	-0.18	-0.17	0.15	0.11	-0.03	0.05	-0.04	0.11	0.23	-0.04	-0.11	-0.12	0.13	-0.07
		ukraine	business suit, news,	0.00	-0.05	0.13	0.06	0.08	0.00	-0.14	0.13	0.07	0.11	0.06	**0.15**	-0.07	-0.02	0.07	-0.03	-0.04	-0.01	0.09
		ukraine	debris, city, rubble,	0.07	-0.05	0.17	-0.08	-0.02	-0.07	0.00	0.00	-0.17	0.00	0.07	-0.05	-0.05	0.18	-0.08	-0.04	-0.17	0.08	-0.10
		covid	interview, sit, table	0.13	0.13	-0.24	-0.13	**-0.29**	-0.14	-0.08	-0.07	0.00	**-0.28**	0.20	0.11	0.01	-0.22	**-0.31**	0.10	0.08	**0.34**	-0.26
		covid	mask, equipment, s	0.08	0.00	**-0.16**	0.02	-0.06	0.08	-0.05	0.05	0.08	0.05	**0.18**	**0.14**	**-0.14**	**-0.10**	-0.06	-0.06	**-0.11**	**0.20**	**-0.14**
		covid	business suit, tie, s	-0.13	0.17	**-0.24**	-0.05	0.07	-0.12	**-0.22**	**0.19**	-0.17	-0.05	**0.20**	0.02	**-0.21**	-0.02	-0.14	0.05	0.00	**0.26**	**-0.26**
	likes	ukraine	camouflage, army,	-0.03	0.15	-0.20	0.11	0.09	0.06	-0.10	0.12	-0.08	0.10	-0.06	0.09	-0.05	-0.13	0.08	0.03	-0.01	0.00	0.04
		ukraine	speech, podium, giv	-0.17	0.45	-0.41	0.28	0.29	0.23	-0.33	0.31	-0.09	-0.24	-0.42	-0.29	0.17	-0.23	-0.08	0.54	0.57	-0.09	0.02
		ukraine	business suit, news,	0.14	**0.17**	**-0.22**	-0.04	**-0.19**	0.06	-0.05	0.06	-0.13	-0.11	0.10	-0.01	-0.01	-0.02	-0.07	-0.01	-0.01	**0.18**	**-0.17**
		ukraine	debris, city, rubble,	0.07	0.15	**-0.22**	0.11	-0.02	-0.10	-0.03	0.03	-0.04	-0.17	0.04	-0.03	-0.03	0.06	0.05	0.13	0.17	0.04	-0.06
		covid	interview, sit, table	0.20	0.17	0.12	**0.34**	-0.16	-0.15	**-0.31**	**0.41**	0.17	0.23	0.05	0.08	-0.03	0.11	0.03	0.19	**0.29**	-0.04	0.08
		covid	mask, equipment, s	-0.08	-0.03	**0.12**	-0.04	0.07	-0.09	**0.19**	**-0.18**	-0.01	-0.07	**-0.13**	**-0.14**	0.08	**0.16**	**0.11**	-0.03	-0.05	0.00	-0.05
		covid	business suit, tie, s	0.12	**-0.23**	0.02	-0.03	-0.12	0.06	0.01	-0.03	0.12	0.00	0.03	0.02	-0.03	0.05	0.03	0.04	0.07	**-0.21**	**0.21**
	comments	ukraine	camouflage, army,	0.08	0.02	0.03	0.06	0.05	0.16	0.04	-0.05	0.07	0.04	-0.11	-0.06	0.09	-0.08	0.02	0.05	-0.01	-0.06	0.02
		ukraine	speech, podium, giv	-0.05	0.49	-0.31	0.43	0.24	0.48	-0.10	0.05	0.09	-0.02	-0.30	-0.03	0.05	-0.47	-0.28	0.34	0.36	-0.29	0.23
		ukraine	business suit, news,	-0.08	-0.04	0.03	-0.07	0.12	0.04	0.10	-0.08	0.06	0.01	-0.01	0.03	0.00	-0.02	0.11	-0.05	0.00	-0.03	0.00
		ukraine	debris, city, rubble,	-0.09	0.15	-0.21	0.06	0.07	0.02	0.13	-0.13	0.07	0.10	-0.04	0.05	-0.07	**-0.25**	0.02	-0.08	0.05	-0.15	0.20
		covid	interview, sit, table	-0.11	-0.12	0.06	**-0.26**	0.19	0.20	-0.08	0.14	-0.05	0.02	-0.03	0.02	-0.12	0.14	0.20	-0.08	0.10	0.02	-0.01
		covid	mask, equipment, s	-0.01	-0.07	0.07	-0.01	-0.06	-0.07	**0.13**	**-0.12**	-0.05	**-0.12**	**-0.29**	**-0.23**	**0.21**	-0.08	**-0.10**	**0.14**	0.04	**-0.17**	**0.10**
		covid	business suit, tie, s	0.01	-0.09	0.06	0.03	0.02	0.07	0.13	-0.11	0.07	-0.04	0.05	-0.04	0.02	0.09	0.16	-0.13	-0.12	-0.16	0.12

Figure 15: Image aesthetics with respect to various video metrics

a separate photograph entirely. Once again, this may be a reflection of Western channel's attempts to use photographs to deliberately create a visually pleasing cinematic scene so as to draw viewers, whereas Chinese channels are more candid in their approach. Another possible explanation for this disparity in symmetry may be that Chinese sources prefer wider shots such as landscapes, where symmetry is easier to achieve because of the lack of micro-details that are captured. This is discussed further later in this section.

Figure 16: Chinese thumbnails with high luminance

The more deliberate choice of thumbnails is also reflected in the Fourier slope feature, as this metric is greater in Western thumbnails, especially in ones relating to 'rubble' and 'debris' in the Ukraine news story. This is because a larger Fourier slope means that high-frequency Fourier features contribute more to the image, ie. there is less of a drop off in power from low-frequency to high-frequency components, and thus finer image details are captured. The increased detail may be a result of higher-grade cameras and the use of proper photography techniques, further supporting the idea that Western thumbnails are more professional photographs whereas Chinese ones are less formal.

Sparseness is more strongly present in Western thumbnails across all events and concepts. This might be because many Western thumbnails are portraits taken with cameras, often at a low aperture causing the background to be slightly blurred. The choice of backgrounds is also often dark backdrops. Both of these factors may make it such that outside of the main subject, there are many homogeneous patches within the image that result in low variance in the activation map whereas Chinese thumbnails don't have blurry background patches giving the image more variance. This may also explain why Chinese thumbnails, on average, score higher in the variability metric than Western ones.

The results from shot scales favor the idea that Western thumbnails have a stronger preference for close-up shots (portraits in most cases) whereas Chinese channels employ wider shots slightly more frequently. In figure 17, across all concepts in the Ukraine news story, Chinese channels show a stronger preference towards long shots than US ones. The results are similar for COVID-19, however, the difference is not as pronounced, especially for certain concepts. In thumbnails of photographing government officials, this may be attributable to the fact that Chinese thumbnails may be extracting frames from

the video itself, where the subject is captured in a medium-long shot whereas Western thumbnails are mostly portraits which are inherently close shots.

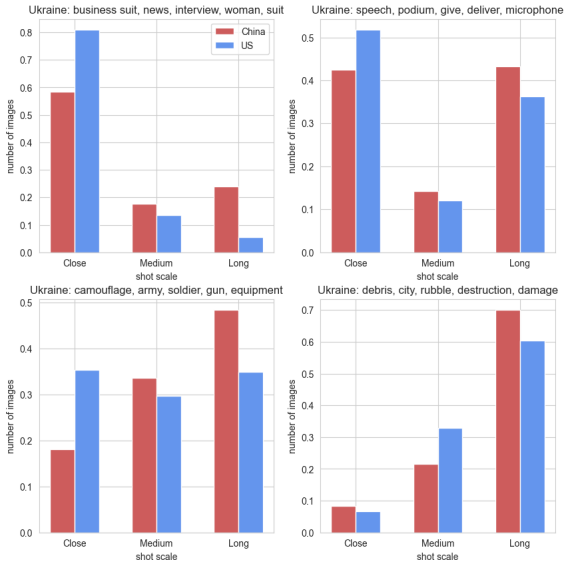

Figure 17: Shot scale distribution for Ukraine concepts

Using bag-of-objects as a descriptor for thumbnails, a comparison of the objects that appear most often in thumbnails can be displayed with a word cloud as in figure 18. Frequent objects, such as a person, tie, or chair, appear in thumbnails from both cultures for both events. Occasionally there are distinct objects, such as a surfboard (in the US, Ukraine category), that only appear once and may suggest errors in object detection. On the aggregated level, there are no detectable cultural differences in our dataset from an object perspective. Analyzing distributional properties, for instance, the relative frequencies of these objects on an image-by-image basis, may reveal subtle disparities.

4.5 Cultural Differences in Thumbnail Aesthetics with respect to Video Performance

The effect of thumbnail aesthetics on different video performance metrics is shown for each culture, event, and concept in figure 15. The correlation of the feature with respect to the corresponding metric is reflected in the cell value and color, where blue means a positive correlation and red means a negative correlation. Whether the relationship is significant or not, i.e. $p < 0.05$ determines whether or not the cell value is bold.

No individual feature alone seems to be an adequate predictor of a video's performance. Each feature, depending on the context of the video (the channel it was uploaded by, the event covered, and the specific concept), can exhibit a positive or negative correlation with each metric. It may be worthwhile to assess the effect of aesthetic features on video metrics in cases where viewer engagement is more important and channels have greater creative freedom over thumbnail graphics such as in entertainment rather than news, a more objective domain.

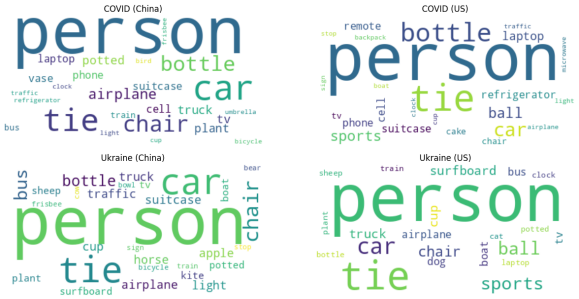

Figure 18: Word clouds of objects from thumbnails

5 Discussion

Our approach allows for a content-controlled analysis of aesthetic features between news images emerging from sources of varying cultural affinity. Using this approach, we find that in the case of YouTube thumbnails relating to COVID-19 and the Ukraine conflict, US news sources prefer to use professional photographs, whereas Chinese media tend to opt for less formal, more candid images, which are likely frames from the videos themselves instead of a separate image altogether. This theme is reflected in many of our observed statistical image properties. Western thumbnails favor more cinematic close-up portrait shots that strongly employ photographic techniques such as the rule-of-thirds and low apertures while appearing to be more edited with features such as deeper color saturation and darker background shadows. In contrast, Chinese thumbnails are often wider shots that are less detailed (weaker presence of high-frequency components), more colorful, and brighter. The above suggests a more natural-looking image as opposed to the more deliberate images from Western sources. Western thumbnails depicting press conferences more frequently set the figure against a dark or natural background, whereas their Chinese counterparts show a strong preference for using blue backdrops. Finally, Chinese thumbnails occasionally express more creative freedom in their use of ani-

mated content, whereas US ones seldom use animated features. Altogether, these may indicate differences in the intentions of Western vs. Chinese sources. YouTube, being more popular in the Western world, is a larger source of revenue for US news platforms, and as such, the selection of thumbnails might be more intentional in order to drive viewer engagement. Dramatizing thumbnails may be reflective of the interests of Western audiences, who potentially view news as some form of entertainment, whereas for Chinese audiences this may be less important. The informal nature of Chinese thumbnails may also be the result of less competition between Chinese channels on YouTube.

Besides image aesthetics, these channels also show some cultural differences in the content they depict in thumbnails. Namely, for the Ukraine conflict, Western sources more strongly emphasize the militaristic narrative, favoring scenes with soldiers and destruction while Chinese sources more frequently cover the geopolitical aspect of the conflict, using thumbnails from conference rooms, speeches, etc. This may reflect the political backgrounds of the sources, as China, being political partners with Russia, who are commonly seen as the aggressors of the conflict, might choose to display the military aspect less so as to not portray Russia in a negative light. Meanwhile, the US, who have openly expressed support for Ukraine, may want to bring awareness to the physical conflict and emphasize the narrative of Ukrainians courageously defending their country. Differences in viewership behaviors between the two audiences are also observed, with Chinese viewers interacting (liking, commenting) with YouTube videos more often than Western ones, which may indicate a tighter community on YouTube or slightly more attentive audiences.

Future investigation may focus on domains where channels have a greater interest in getting users to click on videos and also have more creative freedom over thumbnails, as opposed to a more objective video category such as news. The above factors may cause thumbnail differences to be more pronounced which can serve to emphasize cross-cultural differences. Additionally, focusing on a domain such as entertainment can also lead to increased viewer data (more views, likes, and comments), providing a richer analysis. Although Concept, the topic modeling technique used in this investigation, provided adequate clusters, it is still an open problem other methods for topic modeling news images should be explored. Finding ways to omit channel-specific news watermarks on thumbnails may also benefit an aesthetic comparison as these are often unrelated to the thumbnail itself and may skew the data for certain features. Furthermore, extending this approach to different cultures and news stories may be worthwhile in detecting cross-cultural differences.

References

[1] "Newspapers Fact Sheet." *Pew Research Center's Journalism Project*, Pew Research Center, 10 Nov. 2023, www.pewresearch.org/journalism/fact-sheet/newspapers/.

[2] "Social Media and News Fact Sheet." *Pew Research Center's Journalism Project*, Pew Research Center, 15 Nov. 2023, www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

[3] Smith, Nicholas, and Helene Joffe. "How the public engages with global warming: A Social Representations approach." *Public Understanding of Science*, vol. 22, no. 1, 2012, pp. 16–32, https://doi.org/10.1177/0963662512440913.

[4] Joffe, Hélène. "The power of visual material: Persuasion, emotion and identification." *Diogenes*, vol. 55, no. 1, 2008, pp. 84–93, https://doi.org/10.1177/0392192107087919.

[5] Martikainen, Jari, and Inari Sakki. "How newspaper images position different groups of people in relation to the covid-19 pandemic: A Social Representations approach." *Journal of Community & Applied Social Psychology*, vol. 31, no. 4, 2021, pp. 465–494, https://doi.org/10.1002/casp.2515.

[6] Burgess, Jean, and Joshua Green. *YouTube: Online Video and Participatory Culture*. Polity, 2018.

[7] Xie, Lexing, et al. "Visual memes in social media." *Proceedings of the 19th ACM International Conference on Multimedia*, 2011, https://doi.org/10.1145/2072298.2072307.

[8] Kao, Yueying, et al. "Deep Aesthetic Quality Assessment with Semantic Information." *IEEE Transactions on Image Processing*, vol. 26, no. 3, 2017, pp. 1482–1495, https://doi.org/10.1109/tip.2017.2651399.

[9] Machajdik, Jana, and Allan Hanbury. "Affective image classification using features inspired by psychology and art theory." *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, https://doi.org/10.1145/1873951.1873965.

[10] Anwar, Abbas, et al. *A Survey on Image Aesthetic Assessment*, 7 Feb. 2022.

[11] Lo, Kuo-Yen, et al. "Intelligent photographing interface with on-device aesthetic quality assessment." *Computer Vision - ACCV 2012 Workshops*, 2013, pp. 533–544, https://doi.org/10.1007/978-3-642-37484-5_43.

[12] Redi, Miriam, et al. "The beauty of capturing faces: Rating the quality of Digital Portraits." *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2015, https://doi.org/10.1109/fg.2015.7163086.

[13] Aydin, Tunc Ozan, et al. "Automated Aesthetic Analysis of photographic images." *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, 2015, pp. 31–42, https://doi.org/10.1109/tvcg.2014.2325047.

[14] Datta, Ritendra, et al. "Studying aesthetics in photographic images using a computational approach." *Computer Vision – ECCV 2006*, 2006, pp. 288–301, https://doi.org/10.1007/11744078_23.

[15] Murray, N., et al. "Ava: A large-scale database for aesthetic visual analysis." *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, https://doi.org/10.1109/cvpr.2012.6247954.

[16] Wei Luo, et al. "Content-based photo quality assessment." *2011 International Conference on Computer Vision*, 2011, https://doi.org/10.1109/iccv.2011.6126498.

[17] Bartho, Ralf, et al. *Predicting Beauty, Liking, and Aesthetic Quality: A Comparative Analysis of Image Databases for Visual Aesthetics Research.*

[18] Peng, Yilang, and John B. Jemmott. " Feast for the Eyes: Effects of Food Perceptions and Computer Vision Features on Food Photo Popularity." *International Journal of Communication*, 2018.

[19] Segalin, Crisitina, et al. "The pictures we like are our image: Continuous mapping of Favorite Pictures into self-assessed and attributed personality traits." *IEEE Transactions on Affective Computing*, vol. 8, no. 2, 2017, pp. 268–285, https://doi.org/10.1109/taffc.2016.2516994.

[20] Thomson, T. J., et al. "Politicians, photographers, and a pope." *Journalism Studies*, vol. 19, no. 9, 2017, pp. 1313–1330, https://doi.org/10.1080/1461670x.2016.1268929.

[21] Fahmy, Shahira. "Contrasting visual frames of our times: A framing analysis of english- and Arabic-language press coverage of war and terrorism." *International Communication Gazette*, vol. 72, no. 8, 2010, pp. 695–717, https://doi.org/10.1177/1748048510380801.

[22] Rafiee, Afrooz, et al. "Framing similar issues differently: A cross-cultural discourse analysis of news images." *Social Semiotics*, vol. 33, no. 3, 2021, pp. 515–538, https://doi.org/10.1080/10350330.2021.1900719.

[23] Wallbing, Simon. *Computational Analysis of Swedish Newspapers: Using Topic Detection and Sentiment Analysis*, Feb. 2021.

[24] Chen, Weisi, et al. "Leveraging state-of-the-art topic modeling for news impact analysis on financial markets: A comparative study." *Electronics*, vol. 12, no. 12, 2023, p. 2605, https://doi.org/10.3390/electronics12122605.

[25] Bhargava, Rahul, et al. *Mapping and Visualizing News Images for Media Research*, Mar. 2020.

[26] Zhang, Han, and Yilang Peng. *Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research*, 2021, https://doi.org/10.31235/osf.io/mw57x.

[27] Koh, Byungwan, and Fuquan Cui. "An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content." *Decision Support Systems*, vol. 160, 2022, p. 113820, https://doi.org/10.1016/j.dss.2022.113820.

[28] Kim, Hwiyeon, et al. "Towards visualization thumbnail designs that entice reading data-driven articles." *IEEE Transactions on Visualization and Computer Graphics*, 2023, pp. 1–16, https://doi.org/10.1109/tvcg.2023.3278304.

[29] Chen, Yu-Shih, and John R. Kender. *Differences in Visual Context with Near-Identical Textual Taggings in COVID-19 Videos from China and the US*, 4 Jan. 2022.

[30] "YouTube Data API." *Google*, Google, developers.google.com/youtube/v3/docs. Accessed 4 Jan. 2024.

[31] Zhou, Bolei, et al. *Learning Deep Features for Scene Recognition Using Places Database.*, 2014.

[32] Grootendorst, Maarten P. *Concept*, maartengr.github.io/Concept/index.html. Accessed 4 Jan. 2024.

[33] Huang, Xinyu, et al. *Tag2Text: Guiding Vision-Language Model via Image Tagging*, 2023.

[34] Fadaei, Sadegh. "Comparision of color spaces in DCD-based content-based image retrieval systems." *2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*, 2021, https://doi.org/10.1109/icspis54653.2021.9729360.

[35] Datta, Ritendra, et al. "Studying aesthetics in photographic images using a computational approach." *Computer Vision – ECCV 2006*, 2006, pp. 288–301, https://doi.org/10.1007/11744078_23.

[36] Li, Congcong, and Tsuhan Chen. "Aesthetic Visual Quality Assessment of paintings." *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, 2009, pp. 236–252, https://doi.org/10.1109/jstsp.2009.2015077.

[37] Mallon, Birgit, et al. "Beauty in abstract paintings: Perceptual contrast and statistical properties." *Frontiers in Human Neuroscience*, vol. 8, 2014, https://doi.org/10.3389/fnhum.2014.00161.

[38] Thieleking, Ronja, et al. "Art.pics database: An open access database for art stimuli for experimental research." *Frontiers in Psychology*, vol. 11, 2020, https://doi.org/10.3389/fpsyg.2020.576580.

[39] Schifanella, Rossano, et al. "An image is worth more than a thousand favorites: Surfacing the hidden beauty of Flickr Pictures." *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 9, no. 1, 2021, pp. 397–406, https://doi.org/10.1609/icwsm.v9i1.14612.

[40] Iigaya, Kiyohito, et al. "Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features." *Nature Human Behaviour*, vol. 5, no. 6, 2021, pp. 743–755, https://doi.org/10.1038/s41562-021-01124-6.

[41] Geller, Hannah Alexa, et al. "Statistical image properties predict aesthetic ratings in abstract paintings created by Neural Style Transfer." *Frontiers in Neuroscience*, vol. 16, 2022, https://doi.org/10.3389/fnins.2022.999720.

[42] Peli, Eli. "Contrast in complex images." *Journal of the Optical Society of America A*, vol. 7, no. 10, 1990, p. 2032, https://doi.org/10.1364/josaa.7.002032.

[43] Sidhu, David M., et al. "Prediction of beauty and liking ratings for abstract and representational paintings using subjective and objective measures." *PLOS ONE*, vol. 13, no. 7, 2018, https://doi.org/10.1371/journal.pone.0200431.

[44] Mather, George. "Visual image statistics in the history of Western Art." *Art and Perception*, vol. 6, no. 2–3, 2018, pp. 97–115, https://doi.org/10.1163/22134913-20181092.

[45] Bosch, Anna, et al. "Representing shape with a spatial pyramid kernel." *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 2007, https://doi.org/10.1145/1282280.1282340.

[46] Braun, Julia, et al. "Statistical image properties of print advertisements, visual artworks and images of architecture." *Frontiers in Psychology*, vol. 4, 2013, https://doi.org/10.3389/fpsyg.2013.00808.

[47] Redies, Christoph, et al. "Fractal-like image statistics in visual art: Similarity to natural scenes." *Spatial Vision*, vol. 21, no. 1, 2008, pp. 137–148, https://doi.org/10.1163/156856808782713825.

[48] Brachmann, Anselm, and Christoph Redies. "Using convolutional neural network filters to measure left-right mirror symmetry in images." *Symmetry*, vol. 8, no. 12, 2016, p. 144, https://doi.org/10.3390/sym8120144.

[49] Bertamini, Marco, and Giulia Rampone. "The study of symmetry in empirical aesthetics." *The Oxford Handbook of Empirical Aesthetics*, 2020, pp. 488–509, https://doi.org/10.1093/oxfordhb/9780198824350.013.23.

[50] Brachmann, Anselm, et al. "Using CNN features to better understand what makes visual artworks special." *Frontiers in Psychology*, vol. 8, 2017, https://doi.org/10.3389/fpsyg.2017.00830.

[51] Savardi, Mattia, et al. "Shot scale analysis in movies by Convolutional Neural Networks." *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, https://doi.org/10.1109/icip.2018.8451474.