How Cover Images Represent Video Content: A Case Study of Bilibili

Meredith Dedema
Indiana University, Bloomington
fndedema@iu.edu

Susan C. Herring
Indiana University, Bloomington
herring@indiana.edu

Abstract

User generated videos are the most prevalent online products on social media platforms nowadays. In this context, thumbnails (or cover images) serve the important role of representing the video content and attracting viewers' attention. In this study, we conducted a content analysis of cover images on the Bilibili video-sharing platform, the Chinese counterpart to YouTube, where content creators can upload videos and design their own cover images rather than using automatically generated thumbnails. We extracted four components – snapshot, background, text overlay, and face – that content creators use most often in cover images. We found that the use of different components and their combinations varies in cover images for videos of different duration. The study sheds light on human input into video representation and addresses a gap in the literature, as video thumbnails have previously been studied mainly as the output of automatic generation by algorithms.

Keywords: Bilibili, content analysis, text overlay, thumbnail, video representation.

1. Introduction

Scientists say that our brains can process visual information almost 60,000 times faster than text (McCoy, 2019), which is why our brains love watching videos. With the development of Internet infrastructure and video sharing websites, YouTube has become the most widely used online platforms among U.S. adults (Auxier et al., 2021). When people browse a video website, they expect to see cover images that represent the content of the videos; these help viewers decide whether to watch a specific video. As the phrase goes, "you never get a second chance to make a first impression." The cover image is the viewer's first impression of the video. Therefore, generating the right eye-catching cover image is important to represent the video content and attract viewers. On the YouTube platform, cover images are called "thumbnails" (YouTube Help, 2021). After a video has finished uploading, the YouTube algorithm can automatically generate a thumbnail image that presents a synopsis of the video and that is displayed along with its title. Although the algorithm is good at selecting high-quality images, such as non-blurred images, the generated thumbnails may not necessarily be effective.

While Western audiences are accessing a growing body of user-generated videos online through YouTube, in China, Bilibili is the most popular youth-oriented video streaming and sharing platform, covering 7,000 multi-cultural communities of interest (Tian, 2021). Bilibili's average monthly active users, 86% of whom are below 35 years old, reached 271.7 million in Q4 2021. More than 60% have a university degree, and over half are based in first- and second-tier cities.[1] There are three reasons why Bilibili is thriving among all social media platforms. First, Millennials and Gen Z Chinese consumers pay for content. Virtual gift-giving is a big reason why entertainment livestreaming initially took off in China. With digital payment systems already widely embedded in Chinese apps, users do not have to resort to third-party services (such as Patreon) to support their favorite creators. Second, professional user-generated content on Bilibili accounts for some 90% of content views. Bilibili has high retention in part because it is a place where users have created content to teach just about everything, from coding to cooking. The platform also started streaming educational content during the Covid-19 pandemic. Third, Bilibili offers video content on a variety of topics, including lifestyle (especially vlogs), fashion and beauty, gaming, anime, movies, music, sports, technology, and education (Tian, 2021). Bilibili is also distinctive due to its danmu scrolling comment feature (Teng & Chan, 2022).

The "pay for content" requirement sets a higher standard for the quality of video content. Besides the content itself, the choice of cover image can have a significant impact on a video's views and likes. On the Bilibili platform, the design and use of cover images depend on the video content creators, which means that

[1] https://www.digitalcrew.com.au/blogs-and-insights/bilibili-marketing-guide-2022/, retrieved June 11, 2022

URI: https://hdl.handle.net/10125/102941
978-0-9981331-6-4

the cover images are not generated by the platform or any algorithm but are uploaded by the content creators themselves. Therefore, unlike the YouTube platform where video thumbnails are automatically generated from the video, a cover image on the Bilibili platform can not only be a snapshot from the video, but it can come from other sources. How to design a good cover image is always an important part of video production. Therefore, one might ask, do those user-generated cover images represent the video content well, and how do they represent it? It is expected that short videos (one minute or less) would be easier to represent, as any frame constitutes a larger proportion of the video, compared to longer videos (e.g., more than 20 minutes). Also, since many short videos on TikTok and YouTube Shorts just use the snapshot at the very beginning of the video as the cover image, it might already be a norm to use the first still image for short videos. For long videos, however, content creators might prefer to use a snapshot from the climax of the video or use a more complicated strategy to incorporate as much information in the cover image as possible, so as to attract more viewers.

These issues have never been explored, because video thumbnails have been studied mainly as the output of automatic generation by algorithms. In this study, we address this gap in the literature. We conduct a content analysis of Bilibili cover images and their videos, aiming at understanding the representation of cover images and addressing two specific research questions:

RQ1: What components of cover images represent the video content?

RQ2: How do cover images represent the content in videos of different duration?

2. Literature Review

User-generated content is an emerging force in today's information, communication, and technology industry. The rapid growth of the internet, in terms of both bandwidth and number of users, has pushed all multimedia technology forward, including video streaming (B. Li et al., 2013). As regards Bilibili, for example, studies have been conducted analyzing *danmu* as a new kind of computer-mediated discourse (L. T. Zhang & Cassany, 2020). Scholars aim to understand the characteristics of *danmu* as distinct from comments that appear under the video (Wang, 2021). Y. Zhang et al. (2022) analyzed graphical icons in Bilibili comments over time. Except for the latter study, however, little research has focused on visual affordances of Bilibili. To our knowledge, this study is the first to analyze Bilibili cover images.

The rapid evolution of video content has increased interest in video abstraction techniques (Truong & Venkatesh, 2007). Theoretically, a video abstract could be generated either manually or automatically, but due to the huge volumes of video data and limited human labor capacity, more effort was originally made to develop fully automated video analysis and processing tools for video abstraction (Y. Li et al., 2001).

The word 'thumbnail' originally means "the nail on the thumb." In contemporary parlance, it refers to a very small picture on a computer screen that shows what a larger picture looks like, or what a page of a document will look like when one prints it out. Later, this term was borrowed by video websites such as YouTube to mean a quick snapshot of the video content (W. Liu et al., 2015). In the early stages, the web video thumbnail was seen as one kind of video key frame. Key frame extraction selects some frames from the video, which should represent the contents of the video in the most efficient way (Y. Li et al., 2001). An effective extraction of key frames can greatly facilitate content-based video retrieval, summary, and search, which is defined as the best practice of video summarization in the field of computer vision (Mukherjee & Mukherjee, 2013).

Most conventional methods have focused on learning visual representation purely from video content. The simplest key frame extraction is based on sampling, which randomly extracts a few frames, or extracts frames at specified time intervals (Arman et al., 1994). Shot-based key frame extraction methods that use low-level features such as color and texture could be more effective (T. Liu et al., 2003). In segment-based key frame extraction methods, a video segment could be a scene, an event, or even the entire video sequence. Luo et al. (2009) segmented video clips into homogeneous parts based on major types of camera motion and utilized them to select representative keyframes.

For web video thumbnail generation, more sophisticated methods have been proposed and used. For example, Gao et al. (2009) proposed a theme-based keyframe selection algorithm that explicitly models the visual characteristics of the underlying video theme. Jiang and Zhang (2011) presented a new vector quantization method to create video thumbnails, in which an independent component analysis (ICA) based feature extraction method is employed to explore the spatial characteristics of video frames. Yong et al. (2013) proposed a framework for key-frame extraction, in which the semantic contexts of video frames were extracted and their sequential changes were monitored so that significant novelties were located using a one-class classifier. Song et al. (2016) presented an automatic thumbnail selection system which selects attractive thumbnails by analyzing various objective and subjective metrics (e.g., visual quality and aesthetics) of video frames. They performed clustering analysis to determine the relevance of the video thumbnail to the video content, and they investigated whether selecting a good thumbnail relies highly on objective visual quality metrics, such as frame texture and sharpness.

With more advanced tools, various websites have been created to provide easy-to-use tools to generate GIFs from videos, e.g., GIFSoup, Imgflip, and Ezgif. Gygli et al. (2016) proposed a new model for generating a ranked list of GIFs according to their suitability to highlight and summarize the given videos. From a different perspective, Vasudevan et al. (2017) generated query-dependent thumbnails, aiming at making videos more accessible and searchable via text by identifying the videos' most interesting and representative shots. For each kind of technique for video summarization, video thumbnail generation has been studied mostly by computer scientists and engineering scholars aiming to automatically generate efficient and representative video thumbnails. The selection and the design of thumbnails by video content creators is understudied.

Meanwhile, text overlay in video thumbnails has been overlooked, as most algorithmic methods for thumbnail generation only use video content as input. Yet many sites on the Internet explaining to beginners how to make effective thumbnails for YouTube-like platforms list "include text" as an important component of video thumbnails (Myers, 2020). Text and image can work together to communicate a meaningful message (Kress & van Leeuwen, 2020). Therefore, in this study, we focus on the Bilibili platform as a prime source of manually generated video thumbnails and how they use text overlay, and in so doing, we address the input from content creators themselves to the representation of video content. As video 'thumbnail' always refers to an automatically generated item in prior literature, we use the term 'cover image' to refer to manually generated video thumbnails in this paper. This is to distinguish the terminology, and because 'cover image' is the literal translation of the Chinese expression (封面) used to describe the phenomenon.

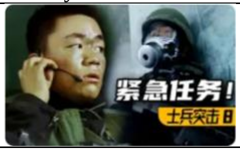
3. Methodology

3.1. Data collection

We used a stratified random sampling approach for collecting the data for this exploratory study, aiming to collect 20 videos for each of four durations: <3 minutes, 3-10 minutes, 10-30 minutes, and >30 minutes. Currently, the Bilibili platform uses <10 minutes, 10-30 minutes, 30-60 minutes, and >60 minutes to support searching for videos of different durations. However, we found that not many videos fall into the >60 minutes category on the platform. Therefore, we broke up the <10 minutes into <3 minutes and 3-10 minutes and combined the 30-60 minutes and >60 minutes into >30 minutes. The reason we investigate videos of different lengths is because the longer the video, the more content it contains; thus, the strategies for representation are likely to be different.

As the Bilibili platform does not provide a random sampling pool for researchers, we used the Weekly Chart to approach random sampling. The Weekly Chart is like the trending function on YouTube; it lists the videos posted in a given week ranked from most popular to least popular, based on certain algorithms developed by the company. We cannot control the topic or any other characteristics of the video with the Weekly Chart, which makes it close to a random sampling procedure. One video was randomly selected from the first page of the Weekly Chart on Bilibili. Subsequently, every 5th result was selected until 80 videos (20 for each duration) were collected. For each video, we collected the video URL, cover image (by screenshot), duration of video, and title of video. See Table 1 for an example.

Table 1. Example of data collection.

ID	#17
Video URL	https://www.bilibili.com/video/BV1XF411h7w7
Duration	>30 minutes
Title	许三多孤身入敌！老 A 全军覆没？《士兵突击》P8 (Translation: Sanduo Xu went to fight the enemy alone! Army A was completely destroyed. *Soldiers Sortie Ep.8*)
Cover Image	

Several judgment criteria were employed in video selection. First, we only included videos in landscape format, where the cover images are horizontally displayed, although we encountered some videos in portrait format; these are mostly for mobile end users (see the example in Figure 1). This variation in the format of cover images and video content could potentially constitute a new dimension for analysis. Second, we excluded videos composed of several independent clips rather than a single stretch of video, because in such cases, the relationship between the cover image and the videos could be more complex.

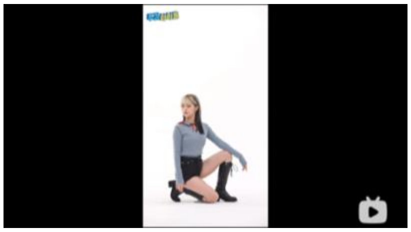


Figure 1. Example of a video in portrait format.

3.2. Data analysis

In the first step, to analyze the data, we used a grounded theory approach (Glaser & Strauss, 1967) to build a codebook (Table 2) based on 12 videos randomly selected from the sample. The coders examined all the components used in the cover images and identified common features. Each feature was defined and operationalized as a code variable through iterative coding. The unit of coding is a pair consisting of a cover image and its video content, and the unit of analysis is the relationship of the cover image to its video content. The variables are snapshot, background, text overlay, and face, which are all important components extracted from the cover image when representing video content. All are binary variables which were coded for presence or absence in the cover image.

Table 2. Codebook.

Component	Description	Value
Snapshot	Does the cover image contain a snapshot from the video?	Yes/no
Background	Does the background of the cover image come from the video (that is, is the background one of the backgrounds shown in the video?	Yes/no
Text overlay	Is there any text overlay in the cover image?	Yes/no
Face	Is any face shown in the cover image?	Yes/no

In the next step, eight videos were randomly selected from the sample to test the codebook. Two coders, both native Chinese speakers, coded these independently. Percentage agreement and Cohen's Kappa were calculated for each variable, with results as follows: snapshot (92%, 0.824), background (92%, 0.833), text overlay (100%, 1), face (100%, 1). This shows a high level of agreement between the two coders. After disagreements were discussed and resolved, all 80 videos were coded as the final sample.

In addition to the thematic analysis introduced above, other characteristics of how each component was used in cover images were considered. We used a qualitative approach when analyzing the cover image and its relationship to video content in order to answer the questions in Table 3.

Table 3. Qualitative analysis.

Component	Questions of interest
Snapshot	Where does the snapshot appear in the video? Is it from the introduction, main body, or the ending of the video?
Background	What are the features of the background? Is it a real setting or a virtual design?
Text overlay	What kind of information is contained in the text overlay?
Face	How many faces are in the cover image? Are the faces present in the video? Are the faces present in the video all shown in the cover image? Whose faces are in the cover image?

4. Findings

4.1. Cover images of videos of different duration

Table 4 shows how cover images use the four components to represent the contents of videos of different durations. As the Bilibili platform does not provide information about the frequency distribution of videos of different durations, we simply add the coding results, rather than weighting them. Results show that *face* (74%) is used most, then *text overlay* (68%) and *background* (45%), while *snapshot* (33%) is least used in cover images. The results of a Chi-square test show that the uses of *snapshot* (χ^2=13.903, p<0.01), *background* (χ^2=12.525, p<0.01), and *text overlay* (χ^2=15.726, p<0.001) differ significantly in cover images of videos of different duration. In contrast, *face* is frequently used in cover images of videos of all durations.

Table 4. Components of cover image in videos of different duration.

	Snapshot		Background		Text overlay		Face		Total	
<3 minutes	13	65%	15	75%	7	33%	15	75%	20	100%
3-10 minutes	4	20%	6	30%	13	65%	13	65%	20	100%
10-30 minutes	6	30%	11	55%	18	90%	16	80%	20	100%
>30 minutes	3	15%	5	25%	16	80%	15	75%	20	100%
Total	26	33%	37	46%	54	68%	59	74%	80	100%

The percentages in Table 4 show that short videos of <3 minutes make use of more snapshots for the cover image (65%).This might be because content creators might not bother to design another cover image when the video is relatively short. This result also nicely echoes the fact that short videos on TikTok and YouTube Shorts typically use the first frame of the video as the cover image. Videos <3 minutes and videos of 10-30 minutes use more background (75% and 50%, respectively) from the video in their cover images. This shows that background is not associated with only short or long videos but rather is used with both, possibly because background from the video is used less as an independent component and more together with other components in cover images. Also, when the cover image uses a snapshot, it guarantees the presence of background, although the opposite is not the case. Long videos use more text overlay in cover images, for example, videos of 10-30 minutes (90%) and videos >30 minutes (80%). This makes sense, because with longer videos it can be challenging to find a single image to represent the entire video's content. Text overlay can help to summarize the content and increase the chance of attracting viewers' attention.

4.2. Use of different components in cover images

4.2.1. Snapshot in cover image. For cover images that use snapshots to represent the video content, the qualitative analysis shows that most snapshots are selected from the introduction of the video, which usually gives an overview or spoiler of the video content. Fewer cover images are selected from the main body or ending of the video. Figure 2 shows an example of a snapshot used as cover image. In 26 videos that use a snapshot as the cover image, only 14 (54%) of them apply text overlay, compared to 68% of videos that use text overlay in general, suggesting that the information contained in the snapshot might already be adequate to represent the video content.

Figure 2. A snapshot used as a cover image.

We also used heat maps to see how much "heat" each snapshot gets in the video. A heat map is automatically generated by the Bilibili platform; it is an analytic tool used to track the number of views on each frame as viewers are watching a video (see example in Figure 3). Heat maps can indicate if content creators are selecting snapshots from the "hottest" part in their videos. In our data, snapshots were mostly selected from higher heat sections (81%). This means that when content creators select snapshots as cover images, they might have a good sense about which snapshot is more representative and is likely to attract more viewer attention. In other words, chances are high that the highlighted frame chosen by the content creators matches the highlighted frame as identified later by viewers.

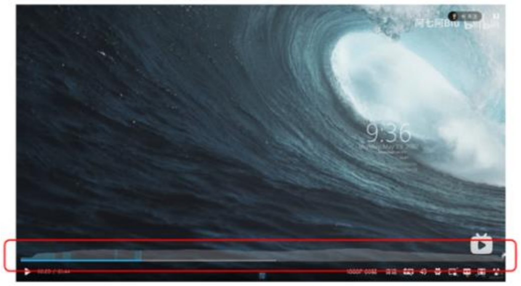


Figure 3. Heat map provided by the platform.

Additionally, there is an interplay between the section that the snapshot is taken from in the video and the heat map. When a snapshot is selected from the introduction section of the video, it is more likely to represent the most viewed period. But this may be because introduction sections naturally get more heat, as people often start watching the beginning of a video but might then discontinue watching it.

4.2.2. Background in cover image. Background is an important component used to represent the video content in cover images. The qualitative analysis shows that 50% of backgrounds use a real setting, while 50% are virtual designs. It is expected that if a cover image is automatically generated by an algorithm based on video content, most, or at least one, of the backgrounds in the cover image should come from the video. But in cover images on the Bilibili platform, since content creators can design the images themselves, more virtual backgrounds are used. This could lead viewers to perceive the videos as more fun, given that more information can be integrated into the cover image in a variety of forms with a virtual design. For example, the cover image in Figure 4 is for a video about how photographers take pictures of the galaxy in the desert. The background in the cover image never appears in the video, but rather the content creator uses it as a virtual background to represent the video content, and it works well.

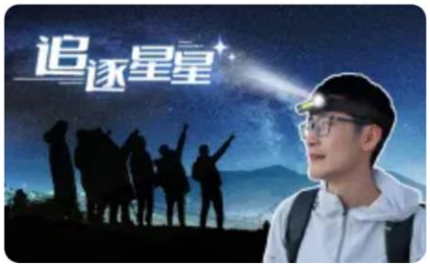


Figure 4. A virtual design background in a cover image.

Overall, only 46% of cover images (see Table 4) use backgrounds from the video, which suggests that identical backgrounds are not seen as effective for representing video content. This might be because the background in videos changes often, making it hard to select just one for a cover image. Among those cover images using background from the video, 70% are real settings, which makes sense, since the videos are usually shot in a real physical setting. When the cover image uses background that is not from the video, chances are higher that the background is a virtual design. Content creators might use a virtual design in a cover image because the real background from the video is cluttered, containing irrelevant objects from the physical world. A virtual background can narrow the focus and cut out distracting information. Also, background is often used together with components like text overlay and face (see Table 8), so it helps to have a simple background when content creators want to emphasize other information in the cover image, such as the title of the video or the people in the video.

4.2.3. Text overlay in cover image. Text overlay is another frequent component used in cover images to represent video content; 68% of cover images use text overlay (see Table 4). Our qualitative analysis asked, "what kind of information is contained in the text overlay?" In response to this question, we developed an exhaustive list, as shown in Table 5.

Table 5. Features of text overlay in cover image.

	Frequency	Percentage
Title	40	74%
Duration	1	2%
Content creator	1	2%
Target audience	1	2%
Other	11	20%
Total	54	100%

Among those cover images using text overlay, 74% present the title of the video. The title of a video contains semantic information, so it can serve as a trailer of sorts, providing hints about the video content (see Figure 5). Two percent of cover images with text overlay mention the duration of the videos, for example, "you only need 3 minutes to learn how to write an essay." This suggests that text overlay can be used not only to convey information about video content but to reflect other characteristics (such as the duration) of the video. Cover images can also use text overlay to convey information about the content creator (2%) to brand or back up the quality of the video, as many content creators already have a reputation based on their prior video production and work. Text overlay also can convey information about the target audience (2%), to attract the attention of a certain group of people, for example, "5 things new hires should know about the workplace." These last examples of text overlay are interesting possibilities for video representation, but they only constitute 6% in total of cover images with text overlay in our sample, suggesting that they are not established norms for text overlay in cover images on Bilibili.

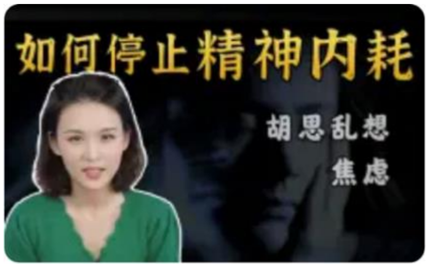


Figure 5. A title used as text overlay in a cover image (Translation: How to stop overthinking and anxiety).

Finally, 20% of text overlays in cover images convey "other" information, including "reference (to video sources)," for example, "breaking news from CGTN"; "channel (of video content)," for example, "come to Mango TV for more fun"; "caption," which are captions or certain lines from the video; and "foreign language," which could be seen as targeting audiences who understand those languages (Figure 6).

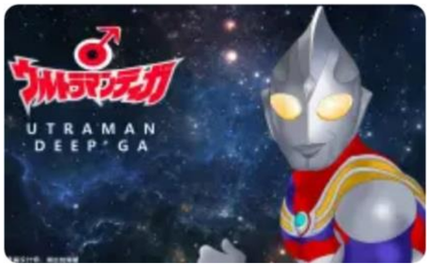


Figure 6. Foreign languages (Japanese and English) used as text overlay in a cover image.

4.2.4. Face in cover image. Face is most often included in cover images, according to the coding results: 59 of 80 videos (74%) include a face in their cover image (see Table 4). We also counted how many faces are shown in the cover image. Table 6 shows that, among those cover images using faces, 58% include only 1 face, 29% include 2-3 faces, 12% include 4-7 faces, and 2% include more than 7 faces. Thus, most cover images show 1-3 faces. We did not systematically analyze the social distance of these faces from the imagined viewer, but most of the faces in our sample appear to be at an intimate or close personal distance (Bell, 2001). This might be because content creators want to show facial expressions with those faces, which could intrigue viewers and encourage them to click on videos and start watching.

Table 6. Number of faces shown in cover image.

	Frequency	Percentage
1 face	34	58%
2-3 faces	17	29%
4-7 faces	7	12%
>7 faces	1	2%
Total	59	100%

As part of the qualitative analysis, we asked, "are the faces present in the video?" For the 59 videos with faces in their cover images, 100% of those faces are present in the videos. It is likely that content creators use actual faces from the video in cover images to represent video content so that viewers will not be confused by finding the faces shown in the cover images but not in the videos. Conversely, we asked, "are the faces presented in the video all shown in the cover image?" Of the 59 cover images with faces, only 29 showed all the faces that were presented in the videos. Displaying faces from the videos in cover images is easy, but displaying all the faces from the videos in cover images can be difficult. Especially if many people appear in the video, it may be impossible or unnecessary to put everyone's face on the cover image. Content creators need to decide which faces are more representative of the video content and which are likely to attract more attention from viewers.

For faces shown in cover images, we also developed a list of whose faces are in the image. As Table 7 shows, faces in cover image are mostly those of the content creator (36%) or a character (36%) from a novel, play, or movie. As Bilibili is mainly a video sharing website for entertainment purposes, many of the videos are recreational or derivative works, in which content creators will use a sample from an original work (e.g., a movie or anime) to generate a video with its own creative storyline. It is common to see characters from movies, TV series, and animation shown in videos and in cover images. Content creators are also shown often in cover images, which is to be expected, given that more and more content creators are producing videos for a living or as a profession. It is important for content creators to brand themselves, and many of them have even become online celebrities. The Bilibili platform has recognized the TOP 100 content creators annually since 2018, to reward content creators' work and influence. After content creators, hosts and guests are next most often shown in the video in their real identities.

Table 7. Types of faces shown in cover image.

	Frequency	Percentage
Content creator	21	36%
Host and guest	10	17%
Character	21	36%
Other	7	12%
Total	59	100%

Faces in the "other" category are either not identifiable or may not be seen as faces by different people. Figure 7 shows an example of two "other" faces. These can be seen as faces since they display eyes, nose, and mouth. But we are unable to find where they are from. They might be from a sticker set, or they could have been drawn by the content creators.

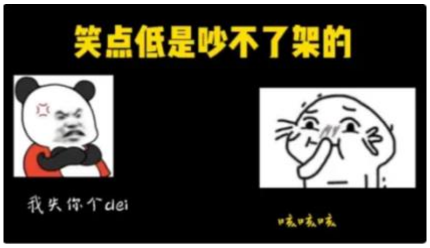


Figure 7. "Other" faces used in a cover image.

4.3. Combination of components in cover images

Snapshot, background, text overlay, face, and their combinations are used in cover images to represent the video content. As Table 8 shows, only 11 (14%) cover images in our sample use all components. The combination of text overlay and face (33%) is used most, which means these two components are most representative of the video when not using a snapshot from the video. No cover images use snapshot only (0%), and few use background only (1%), but some cover images use text overlay only

(8%) or face only (9%), which again underscores the importance of these two components.

Table 8. Components and their combinations in cover images of videos.

	Frequency	Percentage
None	4	5%
Snapshot (only)	0	0%
Background (only)	1	1%
Text overlay (only)	6	8%
Face (only)	7	9%
Snapshot & Background	4	5%
Snapshot & Text overlay	0	0%
Snapshot & Face	0	0%
Background & Text overlay	4	5%
Background & Face	2	3%
Text overlay & Face	26	33%
Snapshot & Background & Text overlay	3	4%
Snapshot & Background & Face	7	9%
Snapshot & Text overlay & Face	0	0%
Background & Text overlay & Face	5	6%
Snapshot & Background & Text overlay & Face	11	14%
Total	80	100%

Finally, there are four cover images in our sample that use none of these components. Figure 8 and Figure 9 are two examples of unrepresentative cover images. Figure 8 shows the workplace of the content creator, where he is responsible for taking care of the cats and dogs in an animal shelter. In this case, the cover image represents the video content well, not by providing information to viewers but by engaging viewers' emotions. We can imagine that viewers will respond "Aww…" when they see this little cat in someone's hand and will click on the video to start watching it with no hesitation. Figure 9 shows a man cosplaying a lovely little girl from Japanese culture (according to the title of the video), but this little girl is not an identifiable character. The cover image only shows a view of "her" back, so when viewers start watching the video, they may be surprised or even shocked. We suspect that is also likely the reaction that the content creator intends to create in his viewers.

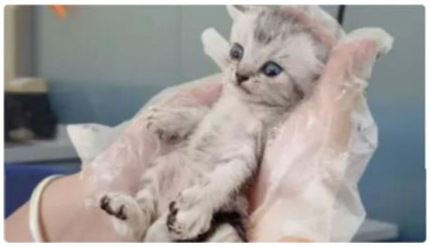


Figure 8. Unrepresentative cover image that evokes compassion.

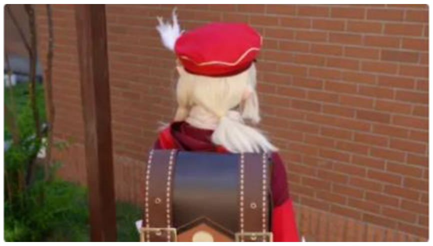


Figure 9. Unrepresentative cover image that evokes curiosity.

In summary, these unrepresentative cover images are well motivated and designed by content creators and are used to elicit a specific kind of reaction from viewers. Instead of representing the video content, these cover images evoke particular feelings in users prior to viewing the video content. This could be an effective strategy to engage viewers' attention, unless some viewers prefer not to start watching a video without understanding the cover image.

5. Discussion

Our first research question asked, *what components of cover images represent the video content*? In other words, how do content creators make connections between the cover images and the video content? Using a grounded theory approach and content analysis, we extracted four components, *snapshot*, *background*, *text overlay*, and *face*. The findings show that *face* and *text overlay* are used most in the cover images in our data, making them more important and representative components compared to *snapshot* and *background*. Backgrounds can be a real physical setting from the video or a virtual background designed by the content creators; each has their advantages. Text overlay often contains information about video content, for example, the title, duration, target audience, or the name of the content creator of the video. Face is most used in cover images as a way to represent people shown in the video and attract attention from the viewers. This aligns with the tips for YouTube beginners, in which face is highlighted as important because the human brain tends to focus naturally on faces (Renderforest, 2021). Putting a human face on a cover image helps create a sense of familiarity and establish a bond with viewers. Those components can mutually interact in a cover image to represent the video content more fully; among these, text overlay and face are most often used together.

The second research question asked, *how do cover images represent the content in videos of different duration*? The findings show that the four components are used differently in cover images for videos of different duration. Short videos use snapshots to represent video content, similar to key frame selection as described in prior literature (Jiang & Zhang, 2011). The difference is that while prior literature is concerned with automatic generation, human content creators select snapshots manually on the Bilibili platform. *Text overlay* is used more in longer videos, since longer videos are harder to summarize with visual cues only. Furthermore, we have observed that the length of a video appears to be related to its genre. Therefore, whether the patterns identified in this paper are manifestations of video length or video genre requires further investigation.

There was also an unexpected finding that some cover images neither use a snapshot nor any of the components identified in this study to represent their video content. This indicates that even though Bilibili content creators generally follow certain norms for video summarization and representation, the selection and design of cover images is still open to possibilities limited only by the human imagination.

6. Conclusions

Cover images or thumbnails are an integral part of video sharing platforms. This exploratory study contributes to their understanding in several respects. It is the first to analyze the relationship between user-generated cover images and their associated video content. As such, the findings shed light on what content creators themselves choose to represent or highlight in cover images. Video thumbnails have previously been studied mainly as the output of automatic generation by algorithms. This study thus addresses a gap in the literature. It shows that human creators' input into cover image production also exhibits patterns, and that snapshot, background, text overlay, and face are used most in cover images for video representation. Importantly, how the different components are used varies for videos of different duration, underscoring the challenges inherent in video representation as regards duration. Finally, the study provides culturally nuanced insights into a major Chinese social media platform.

The results of this study have broader implications for video abstraction. Theoretically, they contribute a better understanding of the relationship between thumbnails and video content, as well as the best means of representing video content. Practically, content creators can learn to use different components for their cover images' design.

Limitations include the small data set, which, although systematically sampled, could limit the generalizability of the findings of the study. Also, the study considers only the case of Bilibili, and the selection and design of cover images could vary on other user-generated video platforms. For the qualitative analysis, although we developed exhaustive lists for each question of interest, it is possible that the values for the information in text overlays and the types of faces in cover image could be more numerous in a larger sample. Also, the small sample size did not allow us to investigate how combinations of different components were used in cover images for videos of different durations. Furthermore, to the best of our knowledge, the cover images were designed by the content creators themselves, yet it is also possible that some used an external AI to generate the cover images. This possibility should be explored in future research.

In future work, it would be worthwhile to interview the content creators and ask directly about their practices in designing cover images. Some automatic analysis of images and videos could also be conducted, for example, examining the color, hue, saturation, and brightness (Reece & Danforth, 2017) of cover images, and examining the camera angle, motion, or change of scene of videos. The relationship between cover images and videos could also be analyzed, along with the number of views, likes, and comments, to examine if more representative cover images attract more attention from viewers. The heatmap results are suggestive in this regard. Finally, a comparative study of two different platforms, Bilibili and YouTube, might usefully be conducted. By uploading the same video to each platform and comparing the cover image automatically generated by the algorithm with the cover image designed by content creators, video thumbnail generation algorithms could be improved with human input. Such a study would also allow us to draw cultural comparisons between media platforms based in Asia and the West.

7. Acknowledgment

We wish to thank Tianji Jiang, who helped with coding and calculating interrater reliability.

8. References

Arman, F., Depommier, R., Hsu, A., & Chiu, M.-Y. (1994). Content-based browsing of video sequences. *Proceedings of the Second ACM International Conference on Multimedia - MULTIMEDIA '94*, 97–103. https://doi.org/10.1145/192593.192630

Auxier, B., & Anderson, M. (2021, April 7). Social media use in 2021. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Gao, Y., Zhang, T., & Xiao, J. (2009). Thematic video thumbnail selection. *2009 16th IEEE International Conference on Image Processing (ICIP)*, 4333–4336. https://doi.org/10.1109/ICIP.2009.5419128

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Aldine Publishing Company.

Gygli, M., Song, Y., & Cao, L. (2016). Video2GIF: Automatic generation of animated GIFs from video. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1001–1009. https://doi.org/10.1109/CVPR.2016.114

Jiang, J., & Zhang, X.-P. (2011). Video thumbnail extraction using video time density function and independent component analysis mixture model. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1417–1420. https://doi.org/10.1109/ICASSP.2011.5946679

Kress, G., & Van Leeuwen, T. (2020). *Reading images: The grammar of visual design*. Routledge. https://doi-org.proxyiub.uits.iu.edu/10.4324/9781003099857

Li, B., Wang, Z., Liu, J., & Zhu, W. (2013). Two decades of internet video streaming: A retrospective view. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *9*(1s), 1–20. https://doi.org/10.1145/2505805

Li, Y., Zhang, T., & Tretter, D. (2001). *An overview of video abstraction techniques* (pp. 1-24). HP Tech Report.

Liu, T., Zhang, H.-J., & Qi, F. (2003). A novel video key-frame-extraction algorithm based on perceived motion energy model. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(10), 1006–1013. https://doi.org/10.1109/TCSVT.2003.816521

Liu, W., Mei, T., Zhang, Y., Che, C., & Luo, J. (2015). Multi-task deep visual-semantic embedding for video thumbnail selection. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3707–3715. https://doi.org/10.1109/CVPR.2015.7298994

Luo, J., Papin, C., & Costello, K. (2009). Towards extracting semantically meaningful key frames from personal video clips: From humans to computers. *IEEE Transactions on Circuits and Systems for Video Technology*, *19*(2), 289–301. https://doi.org/10.1109/TCSVT.2008.2009241

McCoy, E. (2019, February 21). How our brains are hardwired for visual content. *Killer Visual Strategies*. https://killervisualstrategies.com/blog/how-our-brains-are-hardwired-for-visual-content.html

Mukherjee, S., & Mukherjee, D. P. (2013). A design-of-experiment based statistical technique for detection of key-frames. *Multimedia Tools and Applications*, *62*(3), 847–877. https://doi.org/10.1007/s11042-011-0882-2

Myers, L. (2020, June 18). *This is How to Create the Best YouTube Thumbnails*. Louise Myers Visual Social Media. https://louisem.com/198803/how-to-youtube-thumbnails

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, *6*(1), 1–12. https://doi.org/10.1140/epjds/s13688-017-0110-z

Renderforest. (2021). *13 Tips for a Clickable YouTube Video Thumbnail*. https://www.renderforest.com/blog/youtube-video-thumbnail-tips

Song, Y., Redi, M., Vallmitjana, J., & Jaimes, A. (2016). To click or not to click: Automatic selection of beautiful thumbnails from videos. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 659–668. https://doi.org/10.1145/2983323.2983349

Teng, M., & Chan, B. H. S. (2022). Collective colouring in danmu comments on Bilibili. *Discourse, Context & Media*, *45*, 100577. https://doi.org/10.1016/j.dcm.2021.100577

Tian, H. (2021, December). Analysis of the marketing strategy of Bilibili and the reasons for its success. In *2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)* (pp. 2853-2856). Atlantis Press. https://doi.org/10.2991/assehr.k.211209.463

Truong, B. T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *3*(1), 3. https://doi.org/10.1145/1198302.1198305

Wang, J. (2021). How and why people are impolite in danmu? *Internet Pragmatics*, *4*(2), 295-322. https://doi.org/10.1075/ip.00057.wan

Yong, S.-P., Deng, J. D., & Purvis, M. K. (2013). Wildlife video key-frame extraction based on novelty detection in semantic context. *Multimedia Tools and Applications*, *62*(2), 359–376. https://doi.org/10.1007/s11042-011-0902-2

YouTube Help. (2021). *Add video thumbnails*. https://support.google.com/youtube/answer/72431?hl=en

Zhang, L. T., & Cassany, D. (2020). Making sense of danmu: Coherence in massive anonymous chats on Bilibili. com. *Discourse Studies*, *22*(4), 483-502. https://doi.org/ 10.1177/1461445620940051

Zhang, Y., Herring, S. C., & Gan, S. (2022). Graphicon evolution on the Chinese social media platform BiliBili. In *Proceedings of the 4th International Workshop on Emoji Understanding & Applications in Social Media*. https://aiisc.ai/emoji2022/papers/18.pdf