Selecting Representative Thumbnail Image and Video Clip from a Video via Bullet Screen

Yun-Ya Huang, Tong-Yi Kuo, and Hung-Hsuan Chen
ilcilc1975@gmail.com,kuotony860810@gmail.com,hhchen@g.ncu.edu.tw
Computer Science and Information Engineering, National Central University
Taoyuan, Taiwan

ABSTRACT

Finding representative thumbnails or key video clips in a long video usually requires laborious manual editing and compilation. Although there have been deep-learning-based methods that aim to understand videos and pictures, these approaches rely on a large number of computing resources and training data, and the results may still be unsatisfactory. This paper tackles the task of thumbnail selection and highlighting video selection from a different perspective — we leverage on the bullet screen, an emerging new feature on the online video streaming sites that are popular in East Asia, to select thumbnails and video clips. We compared the proposed method with a thumbnail and video clip selecting tool developed by KKStream, a leading streaming service provider in East Asia. We recruited 100 individuals to conduct subjective tests. The experimental results show that most participants are satisfied with the thumbnails and video clips selected by our method, suggesting that the bullet screen could be a valuable resource for video understanding.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Information systems** → *World Wide Web*.

ACM Reference Format:
Yun-Ya Huang, Tong-Yi Kuo, and Hung-Hsuan Chen. 2020. Selecting Representative Thumbnail Image and Video Clip from a Video via Bullet Screen. In *Companion Proceedings of The Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3366424.3382691

1 INTRODUCTION

Making computers understand videos is a challenging task. Recently, researches mostly leverage deep learning techniques to recognize the objects in the key frames in the videos. However, recognization and understanding are different. Geoffery Hinton, known as the father of deep learning, pointed out that although the Convolutional Neural Network (CNN) can identify objects in the image after proper training, the spatial relationship between the objects is difficult to be identified [2], and it is naturally more challenging to understand the images or even the videos.

This paper tackles the video understanding task from a different perspective and uses the representative thumbnail image selection and short video clip selection as the demonstrating scenario. Specifically, we found that certain video streaming websites allow users to comment while watching a video and present the comment by the *bullet screen* format on the video. The bullet screen contains comment text, post time, and, perhaps most importantly, the timestamp, which links the playback time of the video and the bullet screen comments. Traditional comments have the post time but do not have a timestamp, so it is difficult to associate the comment texts with a video segment. Since the bullet screen has timestamp information linking the comment text and the video clip, to some extent the bullet screen comment can be regarded as the metadata of the corresponding video clip, so it is possible to leverage on the bullet screen for video content labeling so the computers may "understand" the videos.

Based on this, we propose a simple yet effective method to select the thumbnail and short clip of a video and name this method *Busk*. Specifically, we found that video clips with a high frequency of bullet screen comments are often the most interesting ones in the videos that cause many discussions among netizens. We crawled bullet screen information from the Internet and recruited 100 individuals to conduct subjective tests on the thumbnails and video clips selected by Busk and by *Stiller*, a thumbnail capturing algorithm developed by KKStream,[1] a leading video streaming service provider in East Asia. It is found that users have positive feedback on the results produced by both methods. However, people who are familiar with the content of the target video (i.e. people who have already watched the video before doing the test) give higher ratings to the results generated by Busk.

2 METHOD

We crawled the bullet screen from 920 thousand videos from the bilibili website.[2] The bullet screen information includes the post time, the comment texts, and the timestamp, i.e., the playback time of the video when the comment is sent.

A large number of bullet screens during a short period may indicate the occurrence of a special or interesting plot. Therefore, we apply an extremely simple yet effective strategy to select the thumbnail image and video clip for a video — we compute the number of bullet screens for every 5 consecutive seconds and pick the one with the most bullet screens. We take extra 5 seconds before the selected clip and extra 5 seconds after the selected clip as the buffer, so eventually, we select 15 consecutive seconds as the video clip to represent the highlight for each video. As for the thumbnail

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3382691

[1] https://www.kkstream.com/

[2] https://www.bilibili.com/

image selection, we pick the keyframe that locates at the center of the selected 15-second video clip.

3 EXPERIMENT

For each video, we generated a thumbnail image and a video clip. We compared the generated results with Stiller [3], an automatic thumbnail selecting algorithm based on the aesthetic clues in the image to attract users clicking and watching the videos. We recruited 100 users to evaluate the thumbnails and video clips selected by Busk and by Stiller as a comparison. We asked each user to select their preference between the results generated by Busk and Stiller. Additionally, we asked each individual to reveal whether they've watched the original video to indicate their familiarity with the original video. Users who are familiar with the video are in Group 1, and the others are in Group 2.

Table 1 and Table 2 show the results. As can be seen, users' average preferences for the representativeness of thumbnails and video clips generated by Busk and Stiller are close. However, if we only focus on the users who are familiar with the original video (i.e., Group 1), these users believe our Busk generates more representative thumbnails and video clips, but users in Group 2 have different opinions. This result is interesting: although Busk may have generated the thumbnails and video clips that better represent the target video (because users who are familiar with the video prefer the results generated by Busk), people who have never seen the video may prefer the thumbnails and video clips selected by Stiller. Therefore, Busk and Stiller seem to be appropriate in different scenarios: if it is to attract new users to watch the video, using Stiller to generate the thumbnails and preview clips may have better results; but if we want to generate the representative thumbnails and video clips, Busk is probably a better choice.

Table 1: Users' evaluation on the representativeness of the outputted video clips

	All users	Group 1	Group 2
Busk	**52.12%**	**67.38%**	47.45%
Stiller	47.88%	32.62%	**52.55%**

Table 2: Users' evaluation on the representativeness of the outputted thumbnail images

	All users	Group 1	Group 2
Busk	47.62%	**63.08%**	43.39%
Stiller	**52.38%**	36.92%	**56.61%**

4 DISCUSSION

To sum up, the timestamp information of the bullet screen allows us to map the scenes in the videos to the text description, and utilizing the frequency of the bullet screen provides a straightforward yet effective method to select the representative thumbnail and video clip for a target video. However, currently, only a few video streaming websites support this unique bullet screen user interface, so the most obvious disadvantage of our method is the limitation of data: Busk cannot be used if there is no corresponding bullet screen of a video. In practice, a more feasible approach may be to use Busk and other keyframe or video clip capturing programs together. However, if the bullet screen information is available, Busk runs much faster, since it requires only simple analysis on the bullet screen. In our experiment, the running time of Stiller is 16% to 54% of the length of the video, roughly 9 minutes to 32 minutes for a one-hour video, whereas our Busk requires only 24 ms to 64 ms.

Finally, although some people have studied the special user interface bullet screen, they mostly focus on the user experience studies [1]. As far as we know, this is the first paper to study the relationship between the bullet screen and the thumbnail or video clip selection. Although we used a simple method, experimental results show that this approach selects the representative thumbnails and video clips, suggesting the emerging feature bullet screen is likely to be a convenient resource for video content labeling. Meanwhile, we already have crawled the bullet screen from over 920, 000 videos, and we are in the process of finding the key video segments and the representative thumbnails based on the texts in the bullet screens. Also, if we can obtain users' IP addresses, we can even observe whether netizens in different regions have different reactions to the same clip. For example, we found that Chinese netizens like the scenes that the Korean football team made mistakes in certain football videos, but the Korean fans may not like these video clips. If we combine the bullet screens with the commenters' IP addresses, we should be able to generate different highlights for the netizens in different locations. Besides, we plan to build a system to recommend video "segments" based on bullet screens. The current video recommendation can only recommend the entire video at a time, but if we use the bullet screen information to create metadata for each segment of the videos, it is possible to recommend related video segments based on the video scene a user is watching. One of our recruited test users said that he is interested in watching the main character of a cartoon speaking up the classic catchphrase in every episode, but current video recommender system cannot accomplish this task, so he is looking forward to the video "segment" recommender system we are currently building.

ACKNOWLEDGMENTS

We appreciate Dr. Jing-Kai Lou at KKStream for his suggestions to this project and Mr. Chun-Ning Tsao for providing the source code of Stiller. We acknowledge partial support by the Ministry of Science and Technology under Grant No.: MOST 107-2221-E-008-077-MY3.

REFERENCES

[1] Soussan Djamasbi, Adrienne Hall-Phillips, Zaozao Liu, Wenting Li, and Jing Bian. 2016. Social viewing, bullet screen, and user experience: a first look. In *2016 49th Hawaii International Conference on System Sciences*. IEEE, 648–657.

[2] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*. 3856–3866.

[3] C. Tsao, J. Lou, and H. Chen. 2019. Thumbnail image selection for VOD services. In *IEEE Conference on Multimedia Information Processing and Retrieval*. 54–59.