Vetrina Attori: Scene Seek Support System Focusing on Characters in a Video

Masahiro Narahara, Kohei Matsumura, Roberto Lopez-Gulliver, Haruo Noma

Vetrina Attori: Scene Seek Support System Focusing on Characters in a Video

Masahiro Narahara[1(✉)], Kohei Matsumura[2], Roberto Lopez-Gulliver[2], and Haruo Noma[2]

[1] Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, 525-8577, Japan
[2] College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, 525-8577, Japan
mnarahara@mxdlab.net

Abstract. In most video services, users watch only the scenes they are interested in, and look back on the scenes they have watched in the past. In these situations, users typically use a seek bar for seeking scenes of the video. They often have to operate the seek bar many times to get to the desired playback position. In this paper, we aim to support the seeking of specific scenes from video contents. In our preliminary study, we found that users seek scenes depending on when each character appears in the video. Therefore, we designed a system to support seeking scenes using the information of characters. We evaluated the usefulness of our proposed system by comparing it with an existing system. According to our qualitative evaluation, we confirm that our proposed system could ease scene seeking.

Keywords: scene seeking · video browsing · character

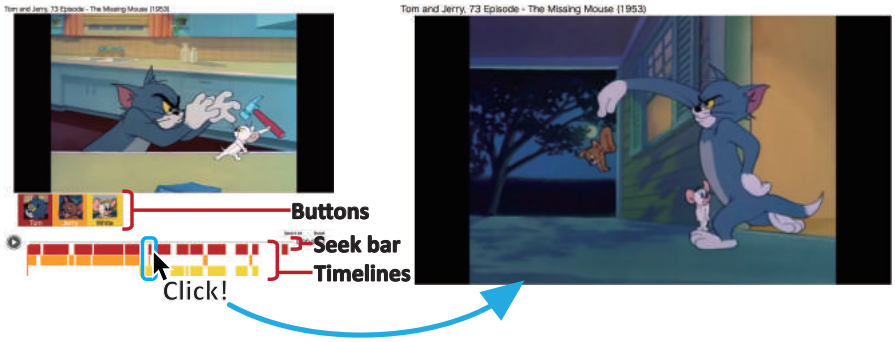

Fig. 1. Interface of our proposed system: Jumping directly to the scene with three characters

1 Introduction

Most Internet users use video services. Video services include video sharing services like YouTube[1] and Niconico[2], and video on-demand services like Netflix[3] and Hulu[4]. The utilization rate of video sharing service reaches 60% to 70% among users [1].

Unlike traditional television broadcasting, the video services have on-demand and interactivity capabilities. On-demand and interactivity allow users to watch video contents at any time and to playback a video from any point of timeline.

Thanks to the on-demand and interactivity capabilities, users enjoy video content in various ways. For example, users watch the same scenes by repeatedly stopping and playing back the video. Some users watch scenes in which a specific character appears in a drama or a movie, or look back on a scene in sports such as baseball. While seeking a specific scene in a video, users usually use a seek bar. The seek bar consists of a slider and a thumbnail. The user can change the playback position of a video with the slider and preview the contents of the playback position with the thumbnail. However, with a seek bar and a thumbnail, it is difficult to set the playback position accurately for a specific scene. The user is required to operate the seek bar many times to adjust the appropriate playback position with the thumbnail. Therefore, we aim to ease scene seeking for video contents by enhancing a traditional seek bar.

There are several methods to support scene seeking in videos. Karrer et al. have conducted studies on Direct Manipulation Video Navigation [2, 3]. In the service Pa-League TV, users watch their own interested play scenes of baseball game [4]. Masuda et al. developed a system that can seek scenes by using annotated tags [5, 6]. However, in these methods, they target specific videos, or they require time and effort from users to annotate them using tags. In this paper, we aim to ease seeking scenes regardless of the video structure or video genres and without requiring extra effort from users.

We firstly conducted a preliminary study and found that participants mainly use the information of characters in scenes while seeking scenes. Based on these findings, we then designed a system that uses the information of characters. Our proposed system generates timelines for each character under an existing seek bar. Fig.1 shows the interface of our proposed system. Users seek scenes by using these timelines as clues.

Finally, we evaluate and verify the usefulness of our proposed system from quantitative and qualitative perspectives. Participants use both the existing system and our proposed system. Results from the quantitative evaluation show that users took more time to seek scenes by using our proposed system. However, results from the qualitative evaluation show that users felt seeking scenes is easier by using our proposed system.

---

[1] https://www.youtube.com/
[2] http://www.nicovideo.jp/
[3] https://www.netflix.com/
[4] http://www.hulu.jp/

2 Proposed System Design and Implementation

2.1 Design

We conducted a preliminary study to investigate what kind of information users use as a clue for seeking scenes in video contents. We gave participants a task that seeking scenes in a video. As the result, we found that users seek scenes using characters as a main clue. Therefore, we use the information of characters to support scene seeking in video contents in our proposed system.

Fig. 1 shows our proposed system. Our proposed system generates timelines, for each character in the video, under a traditional seek bar. Each of the timelines shows the characters' appearance time. By selecting a button with the face and the name of each character, users can switch on or off the display of that character's timeline. The colored part of the timeline shows the time slots that the character appears in the video. When the user selects multiple buttons, timelines of each character are displayed in rows below. As an example of usage (see Fig.1): The video "Tom and Jerry, 73 Episode - The Missing Mouse (1953)[5]" has three characters. If the user wants to watch a scene where all the three characters appear at the same time in the video, the user can find the scene by clicking any place where the three timelines overlap. The background color of each button corresponds to the color of each timeline, and the color is different for each character.

2.2 Implementation

Our proposed system is implemented as follows:

1. The system detects and crops the face area of any character as a face image.
2. The system creates a database that associates face images of the characters with their appearance time in the video.

In our current prototype, we manually cut out the face area of the character as cropped face image and associate the characters with their appearance time. We use these to implement and build the timelines.

First, for each image frame of the video the face area is detected from the face feature points of a person in the image, the face area is cropped from the image frame and saved. We plan to methods for face detection from a video. One is a method using Haar-Like features [7]. The Haar-Like feature characterizes an image by the difference in brightness of the image. Masuda et al. proposed a different method based on Takayama's skin color region extraction [8] to detect the face region of animated characters from videos [9].

Second, we classify the extracted face images for each character, and create a database in which the character and its appearance time are associated with each other using the character information and frame number. This enables us to relate the character and its appearance time.

---

[5] https://www.youtube.com/watch?v=YqGuxOH4Sg4

3 Evaluation

We conducted a series of experiments to evaluate scene seeking time and ease of use of the system. 30 students (20 males, 10 females) aged 21 to 25 participated in the experiment. They regularly use video services such as YouTube and Netflix. In the experiment, we used six genres videos (movie, animation, music, sports, let's play and animal) that are widely shared and distributed in video services.

In the experiment, we asked participants to use the prototype of our proposed system (Fig. 2 left) and the existing system (Fig. 2 right). We replicated the interface of YouTube to reproduce the existing system. Specifically, it has a seek bar to manipulate the playback position of a video and it displays a thumbnail when the pointer is placed on the seek bar.

Fig. 2. Systems used in the evaluation(Left is the prototype of our proposed system. Right is the existing system.)

3.1 Procedure

We explain participants how to use our proposed system and ask them to actually use our proposed system on a test video. After that, we give the participants a task. The task consists of: given a still image from the video as target, seek a scene matching that target image. Participants use our proposed system in three videos chosen randomly and use the existing system in the remaining three videos. Order of videos is randomized and we chose the still image at random. Participants use a laptop computer to watch videos and to seek scenes. We display the still images on the tablet device. The still images are the scenes that another participants chose as interesting scenes in our preliminary study. There are three still images per video. We observe and video record participants during tasks. We ask participants to speak out what they think during the task (think aloud method). After the task, we interview participants.

We evaluate the usefulness of our proposed system from both quantitative and qualitative perspectives. For the quantitative analysis, we compare the difference of seeking time between our proposed system and the existing system. For the qualitative analysis, we conduct a semi-structured interview to participants.

3.2 Results

3.2.1 Quantitative Evaluation

Fig. 3 shows the average scene seeking time using the existing system and our proposed system. In the existing system, the overall average time is 75.0 seconds and the standard deviation is 63.5. In our proposed system, the overall average time is 98.5 seconds and the standard deviation is 96.9. A two-tailed t-test ($\alpha = 0.05$) shows a significant difference in the overall average scene seeking time using the existing system and our proposed system ($p = 0.033 < 0.05$). There is no significant difference in genres other than sports ($p = 0.012 < 0.05$). We think that the reason why the standard deviations are quite large is due to the difficulty of tasks is different depending on still images. Some still images were easy to find and some were difficult to find. From these results, we see that users take more time to seek scenes by using our proposed system.

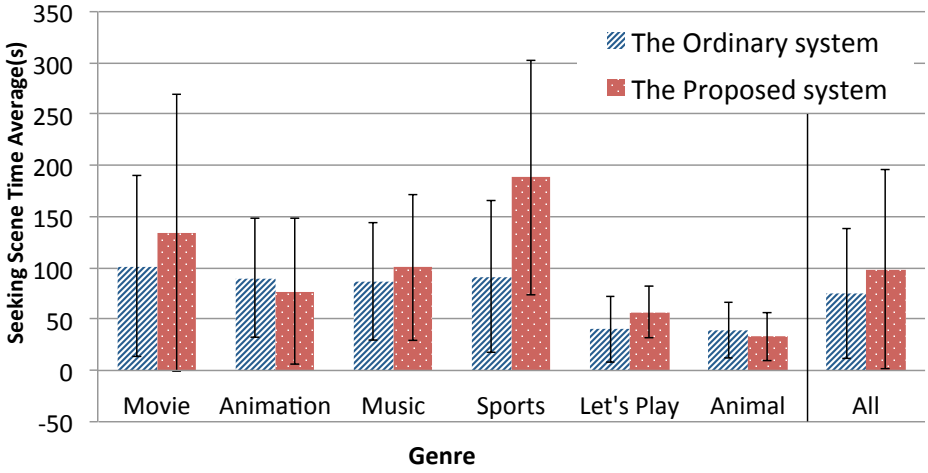

Fig. 3. Results of quantitative evaluation

3.2.2 Qualitative evaluation

Most (26 out of 30) participants reported that seeking scenes feels faster while using our proposed system than the existing one. We found that users felt seeking scenes is faster because our proposed system can reduce the timeline range that they need to seek, as indicated by comments such as "*I could seek a scene easily because the system narrows down the seeking range for me*", "*It is useful because I can find when characters appear and I don't need to seek too many scenes*" and "*I think it is an advantage that I need to seek only the highlighted timeline range.*

Users would prefer using our proposed system when they want to watch scenes that they are interested in, as indicated by comments such as "*I can find scenes easily if characters appear only in specific scenes in a video*", and "*When I want to watch only scenes where my favorite character appears, it is useful.*"

Users would not want to use our proposed system when they watch videos having many similar scenes or when they want to seek scenes with other information, as indicated by comments such as "*If the same characters appear often, there are many similar scenes and it is hard to find the scenes in the timelines*", "*If the same character is appearing from the beginning to the end, eventually I need to watch everything from the beginning*" and "*When I want to seek scenes by using the information of serifs, I can not use this system.*"

In our proposed system, there are still some usage difficulties on the interface and improvement points remain, as indicated by comments such as "*If the same character appears in video from the beginning to the end, I want to seek scenes by using the information of the background rather than characters*", "*Because the face of all characters can be seen as face buttons from the beginning, it spoils users who have not be seen the video before*", "*I wish I could zoom in a specific location of the timeline a bit more finely*" and "*The timelines are so small that I cannot specify the location.*"

4 Conclusion

In this paper, we proposed a system for supporting scene seeking in video contents using information of the characters. From our experiment, we found that users felt faster in seeking scenes while using our proposed system and we also found several problems in our current prototype. In the current prototype, the interface becomes hard to understand when the number of characters is large or the character appears often. We plan to improve the method by changing the way the system displays the characters' buttons or to add the functionality to expand the relevant parts of the timeline.

Our proposed system can visualize the structure of video as timelines. Fig. 4 shows an example in a music video. In the first half of the video, each character appears one by one. After passing the climax, all character appears at the same time in most of scenes. Our proposed system allows users to understand the structure such as verse and climax while playing a music video. There is a possibility that our proposed system can also visualize the structure of other videos as a novel video player.

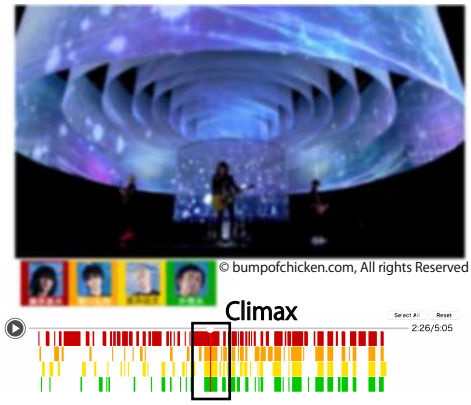

Fig. 4. Example of visualizing the structure of a video

References

1. Ministry of Internal Affairs and Communications, http://www.soumu.go.jp/johotsusintokei/linkdata/h28_02_houkoku.pdf, last accessed 2017/04/10
2. Karrer, T., Weiss, M., Lee, E., & Borchers, J.: DRAGON: a direct manipulation interface for frame-accurate in-scene video navigation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 247-250). ACM. 2008
3. Karrer, T., Wittenhagen, M., & Borchers, J.: DragLocks: handling temporal ambiguities in direct manipulation video navigation. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 623-626). ACM. 2012
4. Pacific-league TV, http://tv.pacificleague.jp/, last accessed 2017/04/10
5. Yamamoto, D., Ohira, S., & Nagao, K.: Weblog-style video annotation and syndication. In: Automated Production of Cross Media Content for Multi-Channel Distribution, 2005. AXMEDIS 2005. First International Conference on (pp. 4-pp). IEEE. 2005.
6. Masuda, T., Yamamoto, D., Ohira, S., & Nagao, K.: Video scene retrieval using online video annotation. In: Annual Conference of the Japanese Society for Artificial Intelligence (pp. 54-62). Springer Berlin Heidelberg. 2007
7. Papageorgiou, C. P., Oren, M., & Poggio, T.: A general framework for object detection. In: Computer vision, 1998. sixth international conference on (pp. 555-562). IEEE.1998
8. Takayama, K., Johan, H., & Nishita, T.: Face detection and face recognition of cartoon characters using feature extraction. In: Image, Electronics and Visual Computing Workshop (p. 48). 2012
9. Masuda, T., Hirai, T., Ohya, H., & Morishima, S.: Recommending Scenes of 2D Characters Based on Image Similarity in Region of Interest. In: Information Processing Society of Japan, 601-602. 2013