An Automatic Video Reinforcing System based on Popularity Rating of Scenes and Level of Detail Controlling

Yuanyuan Wang
Nagoya University, Japan
yuanw@db.ss.is.nagoya-u.ac.jp
Kazutoshi Sumiya
Kwansei Gakuin University, Japan
sumiya@kwansei.ac.jp
Yukiko Kawai
Kyoto Sangyo University, Japan
kawai@cc.kyoto-su.ac.jp
Yoshiharu Ishikawa
Nagoya University, Japan
ishikawa@is.nagoya-u.ac.jp

***Abstract*—With the advance of video-on-demand (VOD) services such as Netfix, users are able to watch many kinds of videos anytime and anywhere. While watching a video, recently, users often search related information about it through the Web by using mobile PC. However, users cannot satisfactorily understand and enjoy it because the video keeps playing when they search about it. It is necessary to detect various questions of the video to supplement their related information about each scene for automatic search. However, only one video includes various topics of each scene, furthermore, viewers have different levels of knowledge. Therefore, we have developed a novel automatic video reinforcing system, called TV-Binder, it generates new video contents from one video stream related to viewers' interests and knowledge by adding other related contents (i.e., YouTube videos, images or maps) and by removing unnecessary original scenes, based on topics of each scene. As a result, viewers can satisfy and joyfully watch modified video contents without searching anything. At first, our system extract topics and detect their scenes of a video stream by using closed captions. The system then searches other necessary contents and determines unwanted original scenes based on popularity rating of each original scene and level of detail (LOD) controlling under time pressure. Through this, TV-Binder can automatically generate video contents are classified into four quadrants by two axes; one is digest and detailed videos, the other one is videos for experts with knowledge about particular topics and ordinary viewers without special knowledge. In this paper, we discuss our automatic video reinforcing system and an evaluation of its effectiveness.**

***Keywords*-topic extraction; scene detection; popularity rating; level of detail (LOD) controlling; closed captions;**

I. Introduction

Recent years have been a huge growth in video-on-demand (VOD) services such as Netfix[1] and Hulu[2], users can watch video contents (e.g., TV and movies) whenever they want to watch and video contents often have closed captions. While watching a video, recently, users often search related information about it through the Web by using mobile PC; they cannot satisfactorily understand and enjoy the video because it keeps playing when they search about it. It is necessary to supplement the video with related information (e.g., web pages, images, or YouTube videos) about each scene for automatic search by detecting various questions of it. However, one video includes various topics of each scene, and viewers have different levels of knowledge. For example, a tourist wants to know foods in Switzerland in a short time, when he watches a video about Switzerland. However, the main topic of this video is history of Switzerland, he cannot watch other related contents of his interests; it is difficult to satisfy the user's requirements by watching only one video in different viewing times. Therefore, it is necessary to extract topics of a video as user interests or questions and to control the viewing time for providing various types of video contents from only one video.

In this work, we aim to develop a novel automatic video reinforcing system, called TV-Binder, to generate new video contents by adding other related contents into a video stream and by removing unnecessary original scenes of this video stream related to viewers' interests and knowledge, based on topics of each scene. As a result, viewers satisfy and joyfully watch modified video contents without searching anything. To achieve our goal, we first extract closed captions of a video stream, and detect topics and their scenes of the video stream. Therefore, our method can measure popularity rating of detected scenes by calculating the number of search hits of topics that appear in each scene. TV-Binder then automatically generate four kinds of new video contents by searching other necessary contents and determining unwanted original scenes, based on a ranking of popularity rating of original scenes and level of detail (LOD) controlling under time pressure. Moreover, we searched additional contents, such as online videos from YouTube[3], images from Google Images[4], or maps from Google Maps[5], based on relevance ranking and viewing time of video contents.

The next section provides an overview of our system and

[1]https://www.netflix.com/
[2]http://www.hulu.com/
[3]https://www.youtube.com/
[4]https://www.google.co.jp/imghp?gws_rd=ssl
[5]https://www.google.co.jp/maps

978-1-5090-0379-2/15 $31.00 © 2015 IEEE
DOI 10.1109/ISM.2015.31

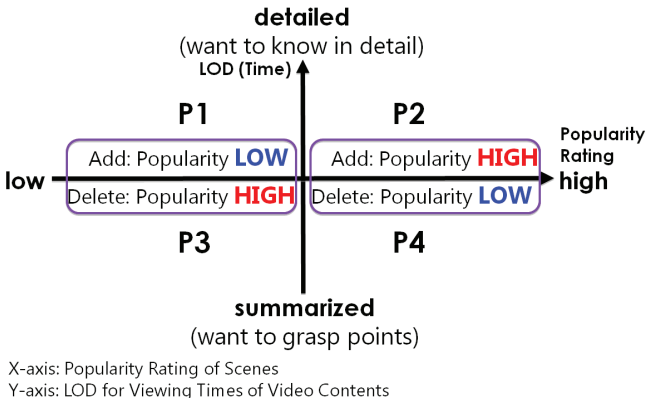

Figure 1. Video contents based on popularity rating and LOD controlling

reviews related work. Section 3 explains how to detect topics and their scenes of video streams. Section 4 describes our research model for generating new video contents. Section 5 discusses experimental results with our prototype system. Finally, Section 6 concludes this paper with future works.

II. System Overview and Related Work

A. Automatic Video Reinforcing System (TV-Binder)

In this work, we propose a novel automatic video reinforcing system called TV-Binder, to generate four kinds of video contents based on popularity rating of each original scene of one video stream and LOD controlling under time pressure. The generated video contents are shown in Figure 1 and they are described as follows:

- **P1**: a detailed video about particular topics for experts
- **P2**: a detailed video about general topics for ordinary viewers (not experts)
- **P3**: a digest video about particular topics for experts
- **P4**: a digest video about general topics for ordinary viewers (not experts)

We classify four quadrants by two axes X and Y to automatically generate video contents **P1**～**P4**; X axis denotes videos for experts who have knowledge about particular topics and ordinary viewers who have no special knowledge by measuring popularity rating of scenes, and Y axis denotes digest and detailed videos by controlling LOD under time pressure. In order to generate **P1** and **P3**, TV-Binder adds other related contents after original scenes that are low popularity rating, and removes original scenes that are high popularity rating. Conversely, in order to generate **P2** and **P4**, TV-Binder adds other related contents after original scenes that are high popularity rating, and removes original scenes that are low popularity rating. In addition, **P1** and **P2** that are detailed videos, then, the viewing time of them may longer than that of the original video stream. Conversely, **P3** and **P4** that are digest videos, then, the viewing time of them may shorter than that of the original video stream. Additional contents such as online videos, images, or maps, from online video sharing sites, image search, or map search, will be added into an original video stream based on high relevance ranking and viewing time of video contents.

An example is shown in Figure 2, which depicts an overview for generating digest video for ordinary viewers (**P4**) by TV-Binder. Scenes with a high or low popularity rating are detected by extracting topics from closed captions of a video stream, and detected scenes are also ranked in an order by their high popularity rating. Red frames denote original scenes with a high popularity rating, and dashed line frames denote additional contents related to topics of original scenes with a high popularity rating, i.e., online videos and images, are added into the original video stream. Meanwhile, blue frames denote original scenes with a low popularity rating, and they are removed from the original video stream. Furthermore, a yellow frame denotes a scene which does not edit in the original video stream.

Based on the above, for instance, scenes of a video stream with a low popularity rating and if you want to gain more information about your interested topics, you can watch a full-length detailed video for experts (**P1**); meanwhile, scenes of a video stream with a high popularity rating and if you want to grasp points related to your interested topics, you can watch a short digest video for ordinary viewers (**P4**).

B. Related Work

Several research efforts have focused on segmenting scenes by clustering of video contents and graph analysis of temporal structures extract from videos [1], [2]. Baraldi et al. [3] divide videos into coherent scenes based on a combination of local image descriptors and temporal clustering techniques. Liu et al. [4] propose a visual based probabilistic framework that detects scenes by learning a scene model. Scene classification in field sport video by using color features and frequency space decompositions [5], [6]. These studies focused on temporal clustering of video contents or visual analysis of color features to divide videos into scenes. Our research aims to automatically generate new video contents from a video stream for satisfying viewers' interests and knowledge at any time of the video stream. Therefore, we extract topics of the video stream related to user interests and knowledge by using closed captions of the video stream, and we detect scenes corresponding to topics.

As in related works about generation of video summaries. Chakraborty et al. [7] develop adaptive summarization techniques, which adapt to the complexity of a video and generate a summary accordingly. Liu et al. [8] proposed a sports video summarization system based on a supervised audio classification that generates the summary video composed of only rally shots. Kawamura et al. [9] summarize sports video automatically using audio and visual information. Meanwhile, many media players allow users to change the playback speed. A technology for controlling the speed of

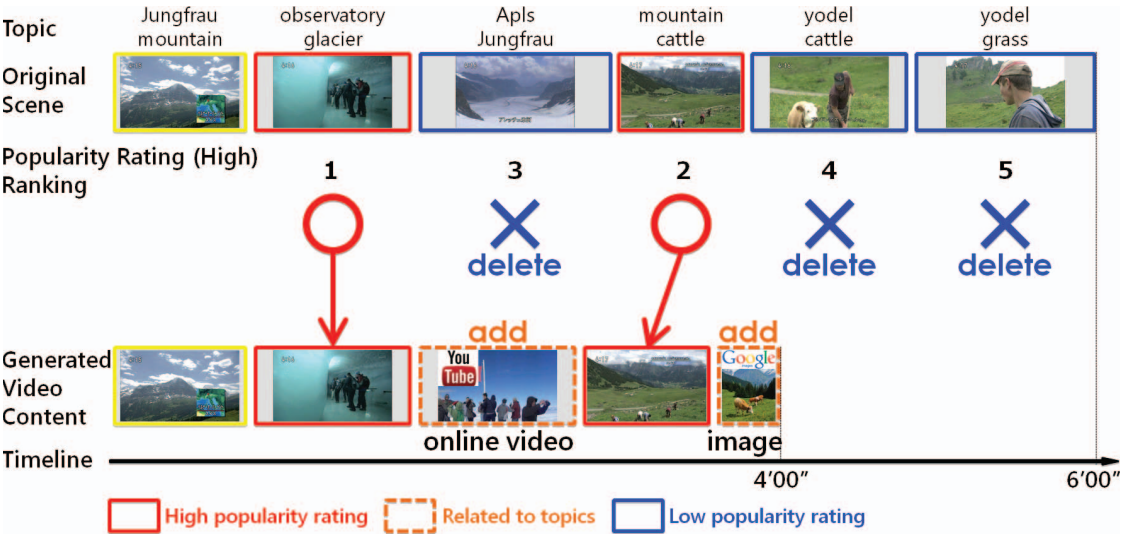

Figure 2. Conceptual diagram of TV-Binder for generating **P4**

playback depending on the context that enables the watching of videos at very high speed, and attaching subtitles that provide useful supplemental information for understanding video contents [10]. Fabro et al. [11] reported a tool for fast nonsequential hierarchical video browsing, which proposed parallel style views for a content. Our TV-Binder is similar to these works, we aim to generate video contents such as short digest videos or full-length detailed videos with LOD controlling by adding other related contents of user interests and by removing unwanted original scenes.

III. Topic Extraction and Scene Detection

In order to extract topics of a video stream, we extract words and their appearance time by using closed captions from a MPEG-2 Transport Stream file of the video stream. Specially, we first extract a bag of words $W_{1,\cdots,i}$ from the closed captions of a video stream by using a morphological analyzer called MeCab[6]. If term frequency tf of one word in $W_{1,\cdots,i}$ exceeds a threshold value, this word will be extracted as a topic. Here, tf returns the term frequency of each word in the closed captions of the video stream, and all topics K of the video stream can be extracted. In this work, one scene that is considered as the unity of content of topics. Therefore, we can divide scenes if the total value of tf of extracted topics along the time sequence of the video stream exceeds than a threshold value α, topics K_j of each detected scene can be acquired.

Figure 3 shows an example of scene detection when α is 0.8. Table columns denote extracted topics and their tf values in descending order from the left of the table; and table rows denote sections of closed captions in an order of time sequence from the top of the table. Therefore, scenes

[6]http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

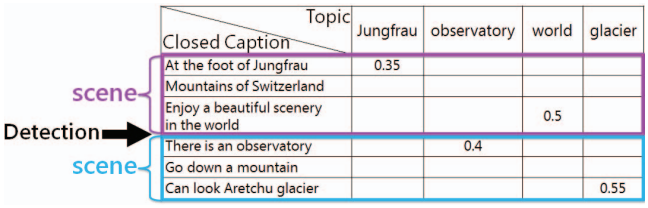

Figure 3. An example of scene detection

can be detected as these two frames when the total value of tf of extracted topics along the time sequence of the video stream is higher than 0.8.

IV. Video Generation based on Popularity Rating of Scenes and LOD Controlling

A. Determination of Adding and Removing Contents

In order to generate new video contents from a video stream, we determine which original scenes should add other related contents, and which original scenes should be removed; the original scenes of the video stream are detected in Section III. First, we use topics K_j of each original scene as a query to search other videos related to each original scene, and we acquire the number of search hits from online video sharing sites such as YouTube. Next, we calculate a threshold value β to determine which original scenes should add other related contents or should be removed by using the number of search hits with the following formula.

$$\beta = \frac{|Search(K)|}{N} \quad (1)$$

Here, $|Search(K)|$ returns the total number of search hits by using all topics K of a video stream. N denotes the total number of detected scenes of the video stream.

Therefore, we measure a popularity rating of each original scene by calculating the number of search hits of each scene $|Search(K_j)|$; and we determine whether add other related contents or remove original scenes by comparing $|Search(K_j)|$ with β as follows:

- **P1** and **P3**
  if $|Search(K_j)| < \beta$ then *low popularity rating, add other related contents*,
  else if $|Search(K_j)| \geq \beta$ then *high popularity rating, remove this original scene*.
- **P2** and **P4**
  if $|Search(K_j)| \geq \beta$ then *high popularity rating, add other related contents*,
  else if $|Search(K_j)| < \beta$ then *low popularity rating, remove this original scene*.

B. Calculation of Viewing Time of Additional Contents

In this work, original scenes should be removed from a high popularity rating ranking to generate video contents (**P1**, **P3**) or a low popularity rating ranking to generate video contents (**P2**, **P4**). In addition, other related contents should be added after original scenes in a low popularity rating ranking to generate video contents (**P1**, **P3**) or a high popularity rating ranking to generate video contents (**P2**, **P4**). Then, original scenes without any modifications in the video stream when they do not satisfy the above conditions. In order to calculate the viewing time t_j of additional contents to control LOD for generating video contents, we calculate a ratio of a popularity rating of each original scene $|Search(K_j)|$ and the total number of original scenes with a low/ high popularity rating N_l/ N_h.

- **P1** and **P3**: $|Search(K_j)| < \beta$

$$t_j = \frac{|Search(K_j)|}{N_l} \times T \tag{2}$$

  Here, N_l denotes the total number of all original scenes with a low popularity rating when their popularity rating is lower than β. T is the total viewing time of additional contents.

- **P2** and **P4**: $|Search(K_j)| \geq \beta$

$$t_j = \frac{|Search(K_j)|}{N_h} \times T \tag{3}$$

  Here, N_h denotes the total number of search hits of all original scenes with a high popularity rating when their popularity rating exceeds β.

In addition, we determine the types of additional contents according to their viewing time by using a threshold value γ based on time pressure with the following conditions.

- if $t_j \geq \gamma$ then *add YouTube videos or Google maps (if topics are location names)*
- if $t_j < \gamma$ then *add Google images or Google maps (if topics are location names)*

Table I
EXPERIMENT I: GENERATED FOUR KINDS OF VIDEO CONTENTS **P1**∼**P4**

	Time	+ Scenes (#)	- Scenes (#)
V1	5'31"	—	—
P1	8'21"	5'05" (3)	2'24" (4)
P2	7'01"	4'45" (2)	3'28" (4)
P3	3'30"	1'07" (3)	3'10" (6)
P4	2'16"	1'00" (2)	4'18" (6)
V2	5'30"	—	—
P1	7'26"	5'17" (5)	3'25" (6)
P2	6'42"	4'51" (3)	3'24" (6)
P3	2'36"	1'40" (3)	4'36" (8)
P4	2'21"	1'00" (1)	4'18" (8)
V3	5'30"	—	—
P1	10'43"	8'27" (6)	3'14" (7)
P2	9'53"	8'11" (3)	3'49" (7)
P3	2'39"	1'34" (3)	4'27" (10)
P4	2'25"	0'47" (3)	4'47" (10)

Table II
EXPERIMENT II: GENERATED FOUR KINDS OF VIDEO CONTENTS **P1**∼**P4**

	Time	+ Scenes (#)	+ Images (#)	+ Maps (#)	- Scenes (#)
V1	5'31"	—	—	—	—
P1	8'21"	5'05" (3)	0'04" (2)	—	2'24" (4)
P2	7'01"	4'45" (2)	—	—	3'28" (4)
P3	3'38"	1'07" (3)	0'08" (2)	—	3'10" (6)
P4	2'24"	1'00" (2)	0'08" (2)	—	4'18" (6)
V2	5'30"	—	—	—	—
P1	7'41"	5'17" (5)	0'16" (4)	—	3'25" (6)
P2	5'37"	4'51" (3)	0'08" (2)	—	3'24" (6)
P3	2'19"	1'40" (3)	0'08" (2)	0'04" (1)	4'36" (8)
P4	3'10"	1'00" (1)	0'12" (2)	—	4'18" (8)
V3	5'30"	—	—	—	—
P1	11'07"	8'27" (6)	0'12" (3)	0'12" (3)	3'14" (7)
P2	10'01"	8'11" (3)	0'04" (1)	0'04" (1)	3'49" (7)
P3	1'19"	1'34" (3)	0'12" (3)	0'04" (1)	4'27" (10)
P4	1'42"	0'47" (3)	0'12" (3)	—	4'47" (10)

Furthermore, we select additional contents based on a low/ high popularity rating of original scenes.

- In order to generate **P1** and **P3**, we select YouTube videos, Google images, or detailed maps from Google Maps based on a high relevance ranking by searching topics of original scenes with a low popularity rating.
- In order to generate **P2** and **P4**, we select YouTube videos, Google images, or extensive maps from Google Maps based on a high relevance ranking by searching topics of original scenes with a high popularity rating.

V. EVALUATION

The purpose of this evaluation with two experiments was to verify whether our proposed TV-Binder is useful for helping viewers to watch their appropriate video contents.

A. Experimental Dataset

As an experimental dataset, TV-Binder generated **P1**∼**P4** from three videos of NHK World Heritage 100[7].

- Original video (V1): Swiss Alps Jungfrau-Aletsch∼ Switzerland∼(viewing time: 5'31")

[7]http://www.nhk.or.jp/sekaiisan/s100/

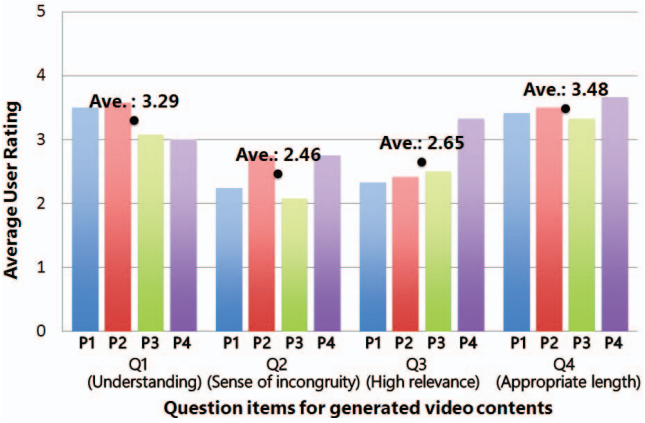

Figure 4. Experiment I: results of $Q1 \sim Q4$ in questionnaire

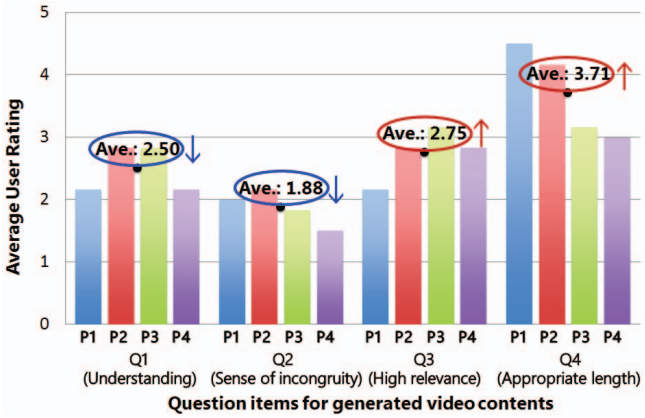

Figure 5. Experiment II: results of $Q1 \sim Q4$ in questionnaire

- Original video (V2): Wachau Cultural Landscape～Austria～(viewing time: 5'30")
- Original video (V3): Yellowstone National Park～U.S. state～(viewing time: 5'30")

Table I shows generated video contents by using only videos as additional contents in Experiment I, and Table II shows generated video contents by using videos, images or maps as additional contents in Experiment II. Here, '+' denotes 'added' and '-' denotes 'deleted.' Furthermore, we determined the types of additional contents according to their viewing time by using a threshold value (γ=10). Overall, our method detected #scenes of V1 was 9, #scenes of V2 was 11, and #scenes of V3 was 13 by using a threshold value (α=0.8); the average viewing time of detected scenes was approximately 30 seconds. Therefore, the average viewing time of detailed videos (**P1**, **P2**) was 8'22", and digest videos (**P3**, **P4**) was 2'32" in two experiments.

There were 10 college students in Kyoto Sangyo University, who participated in the experiments, completed the following 6 items (**Content Understanding**: $Q1$, **Editing Effects**: $Q2 \sim Q4$, **Interest-Arousing**: $Q5$, $Q6$) in questionnaire when they watch generated **P1**～**P4** in experiments I and II, respectively.

- $Q1$: Could understand the video contents.
- $Q2$: No sense of incongruity in switching scenes.
- $Q3$: Unrelated scenes do not exist.
- $Q4$: Felt long about the video content of **P1** or **P2**. Felt short about the video content of **P3** or **P4**.
- $Q5$: Write topics that you interesting in, and how about their interesting levels.
- $Q6$: Write topics that are not related to video contents.

B. Experiment I: Generation by Adding Videos

Figure 4 illustrates the average rating of $Q1 \sim Q4$ in Experiment I by using five-level scales, and high rating may denote good results. Our finding were discussed as follows:

- $Q1$ for detailed videos (**P1**, **P2**) gained a high rating; and $Q1$ for digest videos (**P3**, **P4**) were low rating because several necessary scenes were removed.
- $Q2$ for generated **P1**～**P4** were low rating. Original video streams contain narration, but added videos did not contain narration that participants felt a sense of incongruity in switching scenes. In particular, generated video contents for manias (**P1**, **P3**) were low rating.
- $Q3$ for only digest videos for ordinary people (**P4**) reached a high rating. In particular, detailed videos (**P1**, **P2**) were low rating, since unrelated scenes were added.
- $Q4$ for detailed videos (**P1**, **P2**) that many participants felt long, and digest videos for ordinary people (**P4**) got a highest rating that almost participants felt short. Therefore, we could confirm that viewing time of generated video contents has a good performance.
- $Q5$ for interested topics were written by participants, 80% of them as well as topics are extracted from original video streams for generating **P1**～**P4**. Moreover, it was possible to arouse participants' interests even though several topics that they were not interested before they watching generated video contents.
- In $Q6$, 24 topics were not related to detailed videos (**P1**, **P2**) and 18 topics were not related to digest videos (**P3**, **P4**) by participants. We could find that some topics of them were related to original video streams but added scenes were not appropriate.

In summary, this experiment showed that $Q1$ and $Q4$ for all generated video contents obtained a high rating. $Q2$ for all generated video contents were low rating. In particular, $Q3$ for **P1**, **P2**, **P3** were low rating.

C. Experiment II: Generation by Adding Videos and Images (Maps)

Figure 5 illustrates the average rating of $Q1 \sim Q4$ in Experiment II by using five-level scales, and high rating may

denote good results. The results compared with Experiment I and our finding were summarized as follows:

- $Q1$ for generated **P1** and **P4** were low rating because several necessary scenes were removed as well as Experiment I.
- $Q2$ for generated **P1**∼**P4** were low rating, and the average rating of $Q2$ was lower than that of Experiment I. That is because still images such as images or maps were increased, participants strongly felt a sense of incongruity when they watch the still images.
- Even $Q3$ for detailed videos for connoisseurs (**P1**) were low rating, the average rating of $Q3$ was higher than that of Experiment I. It was considered that related contents became more detailed by using various types of additional contents.
- $Q4$ for detailed videos (**P1**, **P2**) got a high rating that almost participants felt long, and digest videos (**P3**, **P4**) that many participants felt short. Therefore, we could confirm that viewing time of generated video contents has a good performance as well as Experiment I.
- $Q5$ for topics were written by participants that they interested, many topics of them as well as topics are extracted from original video streams. It was possible to arouse participants' interests as Experiment I.
- In $Q6$, 12 topics were not related to detailed videos (**P1**, **P2**) and 8 topics were not related to digest videos (**P3**, **P4**) by participants. Generated video contents except **P4** compared with Experiment I, less topics were not related to original video streams.

In summary, this experiment showed that the average ratings of $Q1$ and $Q2$ for **P1**∼**P4** in Experiment II were lower than those of Experiment I. $Q3$ and $Q6$ for **P1**∼**P4** in Experiment II got good results. The results of $Q4$ and $Q5$ were no clear difference between experiments I and II. In the future, it is necessary to consider sense of incongruity in switching scenes by using other types of additional contents and validity of removed original scenes for generating digest videos. Furthermore, we need to analyze the relevance between additional contents and original video streams.

VI. Conclusion and Future Work

In this paper, we built a novel automatic video reinforcing system, called TV-Binder, it automatically generates four kinds of video contents from a video stream by adding other contents and removing original scenes, based on popularity rating of original scenes and LOD controlling under time pressure. In order to extract topics and their scenes of video streams, we extract closed captions of video streams. Our system then measures popularity rating of scenes by calculating the number of search hits of topics that appear in each scene. We conducted two experiments with our TV-Binder, and we could confirm that the TV-Binder helps users to satisfy and joyfully watch appropriate videos to suit their interests and knowledge levels and watching time.

In the future, we plan to improve the method for selecting and presenting additional contents with other types (e.g, voice, web pages, microblogs), and extract topics by using both closed captions and voice information based on voice recognition. Further, we intend to expand TV-Binder to allow user interactions with the video streams for selecting additional contents and controlling the viewing time.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 26280042, 25280039.

References

[1] M. Yeung, B. Yeo, and B. Liu, "Segmentation of video by clustering and graph analysis," *Computer Vision and Image Understanding*, vol. 71, no. 1, pp. 94–109, 1998.

[2] Y. Song, T. Ogawa, and M. Haseyama, "A scene segmentation approach based on the mcmc method using video structures," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 97, no. 3, pp. 560–573, 2014.

[3] L. Baraldi, C. Grana, and R. Cucchiara, "Scene segmentation using temporal clustering for accessing and re-using broadcast video," in *IEEE ICME 2015*, pp. 1–6.

[4] C. Liu, D. Wang, J. Zhu, and B. Zhang, "Learning a contextual multi-thread model for movie/tv scene segmentation," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 884–897, 2013.

[5] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Transactions on Multimedia*, vol. 7, no. 6, pp. 1097–1105, 2005.

[6] R. Kapela, K. McGuinness, and N. E. O'Connor, "Real-time field sports scene classification using colour and frequency space decompositions," *Journal of Real-Time Image Processing*, pp. 1–13, 2014.

[7] S. Chakraborty, O. Tickoo, and R. Iyer, "Adaptive keyframe selection for video summarization," in *IEEE WACV 2015*, pp. 702–709.

[8] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 415–424, 2009.

[9] S. Kawamura, T. Fukusato, T. Hirai, and S. Morishima, "Efficient video viewing system for racquet sports with automatic summarization focusing on rally scenes," in *ACM SIGGRAPH 2014*, p. 62.

[10] K. Kurihara, "Cinemagazer: a system for watching videos at very high speed," in *ACM AVI 2012*, pp. 108–115.

[11] M. Del Fabro, K. Schoeffmann, and L. Böszörmenyi, "Instant video browsing: a tool for fast non-sequential hierarchical video browsing," *HCI in Work and Learning, Life and Leisure*, pp. 443–446, 2010.