International Journal of Multimedia Data Engineering and Management
Volume 7 • Issue 1 • January-March 2016

An Automatic Video Reinforcing System for TV Programs using Semantic Metadata from Closed Captions

Yuanyuan Wang, Yamaguchi University, Ube, Japan
Daisuke Kitayama, Kogakuin University, Tokyo, Japan
Yukiko Kawai, Kyoto Sangyo University, Kyoto, Japan
Kazutoshi Sumiya, Kwansei Gakuin University, Sanda, Japan
Yoshiharu Ishikawa, Nagoya University, Nagoya, Japan

ABSTRACT

There are various TV programs such as travel and educational programs. While watching TV programs, viewers often search related information about the programs through the Web. Nevertheless, as TV programs keep playing, viewers possibly miss some important scenes when searching the Web. As a result, their enjoyment would be spoiled. Another problem is that there are various topics in each scene of a video, and viewers usually have different levels of knowledge. Thus, it is important to detect topics in videos and supplement videos with related information automatically. In this paper, the authors propose a novel automatic video reinforcing system with two functions: (1) a media synchronization mechanism, which presents supplementary information synchronized with videos, in order to enable viewers to effectively understand the geographic data in videos; (2) a video reconstruction mechanism, which generates new video contents based on viewers' interests and knowledge by adding and removing scenes, in order to enable viewers to enjoy the generated videos without additional search.

KEYWORDS

Geographical Metadata, Geographical Relationships, Media Synchronization Mechanism, Popularity Rating, Scene Detection, Topic Extraction, Topical Metadata, Video Reconstruction Mechanism

DOI: 10.4018/IJMDEM.2016010101

INTRODUCTION

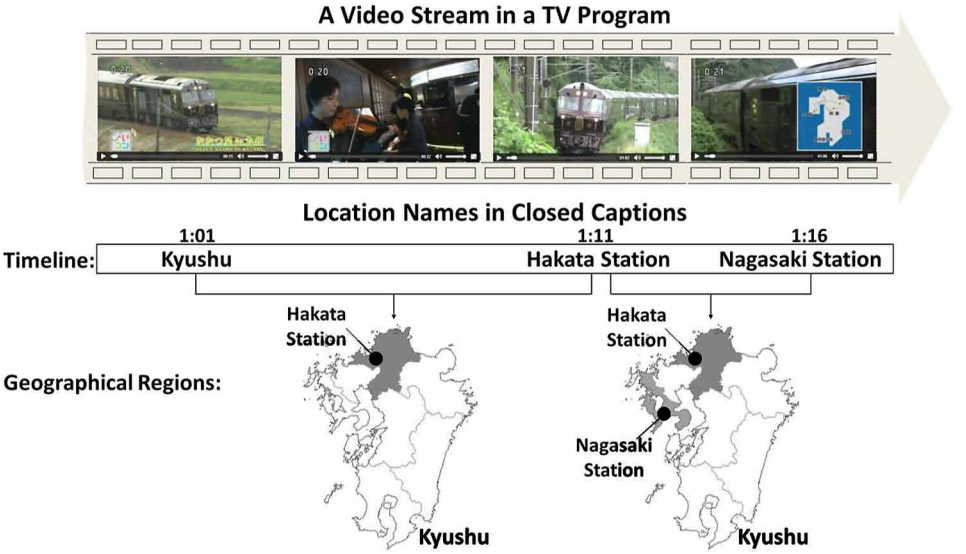
In recent years, there has been a rapid growth in TV channels all over the world, such as CBS (Columbia Broadcasting System) in the United States and NHK (Nippon Hoso Kyokai) in Japan. There are usually many kinds of TV programs (e.g., TV shows, news, and sports events), and the videos in TV programs are often associated with closed captions. While watching TV programs, viewers are probably interested in some contents in the videos and search related information through the Web. For example, viewers may search locations of tourist spots appeared on a travel channel, check the information of a player in a live sports program, or access an online store in a fashion program (Fleites, Wang, & Chen, 2015a, 2015b) using smartphones or tablets. However, since TV programs keep playing and cannot be paused, when searching the Web viewers possibly miss some important scenes and their enjoyment would be spoiled. Another problem stems from the various topics of each scene in a video and the different levels of knowledge of viewers. For example, some tourists may want a summary of delicious foods in Switzerland, while others may be more interested in the historical information about Switzerland. In other words, there can be different demands when viewers watch the same TV program about world heritage sites in Switzerland. On the other hand, they have to refer to other TV programs or resources for the wanted information such as Swiss foods or the history of Switzerland. In other words, it is difficult to meet multiple demands of viewers within only one video. Therefore, it is necessary to extract various topics from the video, which serve as viewers' interests, and supplement the video automatically with related information (e.g., geographic contents, web contents) in each scene.

In this work, the goal is to develop a novel automatic video reinforcing system by analyzing semantic metadata (geographical metadata and topical metadata) from closed captions of videos in TV programs. The proposed system includes two functions: (1) a media synchronization mechanism and (2) a video reconstruction mechanism described as follows. This work is extended from the existing work (Wang, Kawai, Sumiya, & Ishikawa, 2015) which includes only the video reconstruction mechanism. The authors develop the media synchronization mechanism based on the concept of second screen service (Nandakumar & Murray, 2014; Geerts, Leenheer, Grooff, Negenman, & Heijstraten, 2014):

1. **A media synchronization mechanism:** This mechanism presents additional geographic contents (e.g., map, Street View) synchronized with videos on large size screens or smaller sub-screens (smartphones) based on geographical relationships between every two location names appeared in each scene. To achieve it, the system extracts geographical metadata of each video, including a) temporal sequences of location names appeared in closed captions of the video; b) geographical relationships between the locations in each scene based on their geographical regions (see Figure 1). Then, the authors obtain semantic structure such as the intentions of the video. Finally, the authors determine how to present geographic contents seamlessly during the video;
2. **A video reconstruction mechanism:** This mechanism integrates other related web contents (e.g., YouTube video clips, images) into a video and generates new video contents based on viewers' interests and knowledge. Also, it removes unnecessary original scenes based on popularity rating of each original scene on related topics. To achieve it, the system extracts topical metadata of a video, including a) temporal sequences of topics appeared in closed captions of the video; b) popularity rating of each scene based on the number of search hits on related topics. Then, the authors determine other necessary contents and unnecessary original scenes. Finally, the authors determine how to generate four kinds of new video contents with level of detail (LOD) controlled under time pressure.

The proposed novel system enables viewers to obtain geographical information (e.g., tourist spots, routes) easily and effectively during a video by the media synchronization mechanism, and enjoy the video with modified video contents by the video reconstruction mechanism without additional search.

Figure 1. Geographical metadata of a video in a TV program



The remainder of this paper is structured as follows. Firstly, the authors provide an overview of the proposed system and review related work. Secondly, the authors explain the research model by analyzing semantic metadata of videos. Thirdly, the authors describe the functions of the geographical content synchronization and the video reconstruction, respectively. After these sections, the authors discuss experimental results of the prototype system. Finally, the authors conclude this paper and present future works.

SYSTEM OVERVIEW AND RELATED WORK

In this section, the authors describe the proposed automatic video reinforcing system, including a media synchronization mechanism and a video reconstruction mechanism. In addition, the authors briefly review related work on utilizing geographical information for TV programs and generating video contents.

Automatic Video Reinforcing System

Media Synchronization Mechanism

Most travel and history programs contain dense geographic data, such as tourist spots or historical places. However, viewers may find it difficult to grasp the surroundings of these spots or places, understand how the locations are related, and know distances between them, when hearing them instantaneously in the videos. The proposed system synchronizes geographic contents with TV programs on the basis of the derived semantic structure of videos, which aids viewers to better grasp the geographical information (e.g., tourist spots, regions, routes) presented. In a large size screen, a television map interface is implemented using JavaScript by four parts: a video player window (left top), a list of location names (left bottom), a map window (center bottom), and a Street View (photo) window (right) (see Figure 2). When playing a video in the video player window, geographical metadata is automatically extracted from closed captions by using Scala. Thus, the user interface can show the video synchronized with a map using Google Maps API (Google Maps, 2016) in a map window, and Street View using Google Street View Image API (Google Street View Imagery, 2016) or online photos using Panoramio API (Panoramio, 2016) in the Street View (photo) window.

Figure 2. A television map interface of the media synchronization mechanism

In addition, a map or a Street View (photo) can be interactively presented on second screen devices such as tablets and smartphones.

With the media synchronization mechanism, when two location names appear in closed captions of one scene in a video, a route between these two locations can be shown to viewers with the video by using Street View. In addition, in order to grasp the surroundings and positions of the locations and distances between them, the interface presents not only Street View, but also Google Maps or online photos. Therefore, the media synchronization mechanism can enable viewers to effectively grasp the geographical information of the locations, which appear but are not described in the original videos, especially in travel programs.

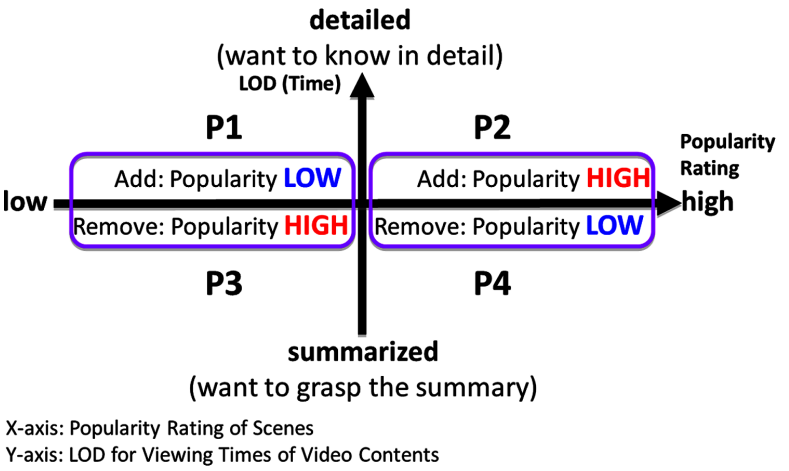
Video Reconstruction Mechanism

In addition to locations, there are various topics appeared in many TV programs, such as sports and educational programs. Viewers usually have different topic interests and different levels of knowledge. The system generates four kinds of video contents from a video based on popularity rating of each original scene with level of detail (LOD) controlled under time pressure. This will help different viewers to enjoy video contents that suit their interests and knowledge levels. The generated video contents from a video are shown in Figure 3 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 (non-experts).
P3: A digest video about particular topics for experts.
P4: A digest video about general topics for ordinary viewers (non-experts).

They are classified into four quadrants by two axes X and Y, where X axis denotes videos for experts who have knowledge about particular topics and ordinary viewers who have no special knowledge (non-experts) by measuring popularity rating of scenes, and Y axis denotes detailed videos and digest videos by controlling LOD under time pressure. In order to generate **P1** and **P3** for experts, the mechanism adds other related contents after original scenes that are with low popularity ratings, and removes original scenes that are with high popularity ratings. On the contrary, in order to generate **P2** and **P4** for non-experts, the mechanism adds other related contents after original scenes

Figure 3. Four kinds of video contents based on popularity rating and LOD controlling



that are with high popularity ratings, and removes original scenes that are with low popularity ratings. In addition, the viewing time of detailed videos (**P1**, **P2**) may be longer than that of the original video, and the viewing time of digest videos (**P3**, **P4**) may be shorter than that of the original video. Moreover, additional contents such as online videos are acquired from YouTube. And images are obtained from Google Images and maps are acquired from Google Maps based on relevance ranking and viewing time of generated video contents.

Related Work

Utilization of Geographical Information

Many studies have focused on the extraction and utilization of geographical information from various contents. Takahashi et al. (2010) proposed a ranking method applied to earth science data. They extracted temporal information and geographical information from articles based on the link structure of Wikipedia. Kitayama et al. (2010) proposed a method to enhance a digital map interface by reflecting the users' intentions with automatically customized visible objects on maps. They determined the types of the objects based on user's operations and relations of the object's appearing patterns between location names. In this work, although the proposed media synchronization mechanism is similar to these works, the system extracts and utilizes geographical information of TV programs, and presents supplemental geographic contents for understanding the TV programs.

On the other hand, some studies have tried to provide location-aware interface by exploring the geographical information to help users in terms of localization. Viana et al. (2011) proposed a method using context awareness and semantic technologies in order to improve and facilitate the organization, annotation, retrieval and sharing of personal mobile multimedia documents. This method integrates metadata extracted from the user context (e.g., locations, geographical relationships). Chi et al. (2009) proposed a model as a guideline to develop mobile applications for tourism. They are similar to this work in the way that the authors propose a television map interface for location awareness by extracting geographical metadata, i.e., location names and geographical relationships.

As in related works about the generation of various geographical contents according to the purposes of users, Banjou et al. (1997) proposed a method for generating rough maps according to the purposes by interactive trial and error with user requirements. Haklay & Weber (2008) proposed a method for providing user-generated street maps in a knowledge collective project that follow the peer production model created from Wikipedia. Evers et al. (2007) described an experiment that aims at determining the most effective and natural orientation of a virtual guide to give route directions

in a 3D virtual environment. Kobayashi et al. (2011) proposed a system for transforming a modified map into a streaming video based on intentions of a map's maker. In this work, the authors propose an interface for presenting geographic contents synchronized with TV programs, which utilizes geographical information in videos by considering the semantic structure of the videos.

Generation of Video Contents

Over the past few decades, a considerable number of studies have been conducted on the generation of video summaries. Chakraborty et al. (2015) developed adaptive summarization techniques, which adapt to the complexity of a video and generate a summary accordingly. Liu et al. (2009) 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. (2014) summarize sports video automatically using audio and visual information. On the other hand, many media players allow users to change the playback speed. A technology for controlling the speed of playback depending on the context that enables users to watch videos at very high speed, and attaching subtitles that provide useful supplemental information for understanding video contents has also been proposed (Kurihara, 2012). Foulke et al. (1969) reported the SOLAFS algorithm for changing the speed of speech without pitch shifting. They concluded that it was effective for rapid understanding of content. Fabro et al. (2010) developed a tool for fast non-sequential hierarchical video browsing, and proposed parallel style views for a content. In this work, while the proposed video reconstruction mechanism is similar to these works, the authors aim to generate video contents such as short digest videos or full-length detailed videos with LOD controlling by adding other related contents based on users' interests and removing unwanted original scenes.

For reconstructing and generating video contents, it is necessary to detect scenes from videos. Video decomposition techniques aim at partitioning a video into sequences, like shots or scenes, according to semantic or structural criteria. Several research efforts have focused on segmenting scenes by clustering of videos and graph analysis of temporal structures extracted from videos (Yeung, Yeo, & Liu, 1998; Song, Ogawa, & Haseyama, 2014). Baraldi et al. (2015) proposed a method for dividing videos into coherent scenes based on a combination of local image descriptors and temporal clustering techniques. Liu et al. (2013) proposed a visual based probabilistic framework that detects scenes by learning a scene model. To solve a problem of efficient description of the video scene sequence, several techniques have been developed for scene classification in field sports videos by using color features and frequency space decompositions (Rasheed & Shah, 2005; Kapela, McGuinness, & O'Connor, 2014). These studies focused on temporal clustering of video contents or visual analysis of color features in order to divide videos into scenes. In this work, the authors aim to automatically generate new video contents from a video for satisfying viewers' interests and knowledge at any time of the video. Therefore, the video reconstruction mechanism extracts topics of a video related to users' interests and knowledge by using closed captions of the video, and detects scenes corresponding to the extracted topics.

SEMANTIC METADATA ANALYSIS FOR TV PROGRAMS

In this section, the authors provide the details of the proposed research model to analyze semantic metadata, i.e., geographical metadata and topical metadata of videos. The authors first extract a bag of words, i.e., location names or topics, from closed captions of a video by analyzing MPEG-2 Transport Stream files (with the HbbTV standard) of the video.

Geographical Metadata Extraction

The geographical metadata can be extracted by detecting location names and their scenes, and extracting geographical relationships between the locations in each scene.

Detecting Location Names and Scenes

The system extracts location names using a morphological analyzer and the temporal sequence, which is the order of appearance of the location names in a video. In the media synchronization mechanism, the authors need to detect scenes according to the geographic data and the time width of each scene. For this purpose, the system converts all location names into latitude and longitude coordinates, and obtains the latitude and longitude coordinates (X, Y) of a center point c of all locations by the following formula:

$$(X,Y)=\left(\frac{\sum_{j=1}^{n}x_j}{n},\frac{\sum_{j=1}^{n}y_j}{n}\right) \tag{1}$$

Here, the authors assume that there are n location names appeared in closed captions of the video. x_j denotes the latitude coordinate and y_j denotes the longitude coordinate of the location j.

Therefore, the authors determine one scene of the video by considering both geographical distance and temporal distance between every two locations in their appearance order. In addition, each location has its own region, such as regions of prefectures, municipalities, and countries. Suppose that the region of each location is represented as a circle. The authors need to measure the geographical distance between two locations by considering inclusion relationships between their regions and the radius of each region. The determination of one scene is described as follows:

$$\begin{cases} distG(j,j+1)-r(j)-r(j+1) & \leq \quad \alpha \\ distT(j,j+1) & \leq \quad \beta \end{cases} \tag{2}$$

$$\alpha=\frac{\sum_{j=1}^{n}\left|distG(c,j)-r(j)\right|}{n-1} \tag{3}$$

$$\beta=\frac{\sum_{j=1}^{n-1}distT(j,j+1)}{n-1} \tag{4}$$

Function r returns the radius of the region of a location. $r(j)$ denotes the radius of the location j. Function $distG(j, j+1)$ returns the geographical distance between the center coordinates of two locations j and $j+1$ in their appearance order. $distG(j, j+1)$-$r(j)$-$r(j+1)$ denotes the geographical distance between adjacent boundaries of the regions of j and $j+1$. The threshold value α in Equation (3) is the average geographical distance between the center point c of all locations in Equation (1) and the adjacent boundary of the region of each location. Function $distT(j, j+1)$ returns a time width, which is the temporal distance between j and $j+1$ in their appearance order. The threshold value β in Equation (4) is the average temporal distance between every two locations. If two locations satisfy the above conditions, one scene is determined by these two locations.

Extracting Geographical Relationships

Since each location has its own region, such as the regions of prefectures, municipalities, and countries, the system extracts geographical relationships, specifically regional relationships, between every two

locations along the timeline in each scene of a video. In this work, the authors utilize 9-intersection (Kurata, 2010) to extract regional relationships between two locations (see Figure 4).

As shown in Table 1, the authors extract six patterns of geographical relationships between two locations. The authors unify *cover* into *contain*, and unify *coveredBy* into *inside*. *Equal* means that the same location name appears one time or repetitively. For example, Kyushu appears continuously in a video. *Disjoint* means the regions of two locations are separated, such as Fukuoka Prefecture and Kagoshima Prefecture in Japan. *Meet* means the regions of two locations have an overlapping boundary, e.g., Fukuoka and Saga Prefectures in Japan. *Overlap* means the regions of two locations are overlapping, e.g., Hitoyoshi City and Kuma River in Japan. *Contain* means the region of one location includes the region of another location in the video such as Fukuoka Prefecture and Hakata Station in Japan. *Inside* means the region of one location is included in the region of another location in the video, e.g., Hakata Station and Fukuoka Prefecture in Japan.

Topical Metadata Extraction

Detecting Topics and Scenes

The system also extracts topics using a morphological analyzer and the temporal sequence, which is the order of appearance of the topics of a video. If the *tf-idf* value of one word in each section of closed captions exceeds the average *tf-idf* value of all words, this word will be extracted as a topic. In this way, all topics K of the video can be extracted. In the video reconstruction mechanism, the authors consider that one scene is the collection of topics. The authors then divide scenes if the total *tf-idf* value of extracted topics along the timeline of the video, exceeds a threshold value γ. In this way, the topics K_j of each detected scene j can be acquired. γ is defined as the average *tf-idf* value of all words in each section of closed captions.

Figure 4. Regional relationships based on 9-intersection

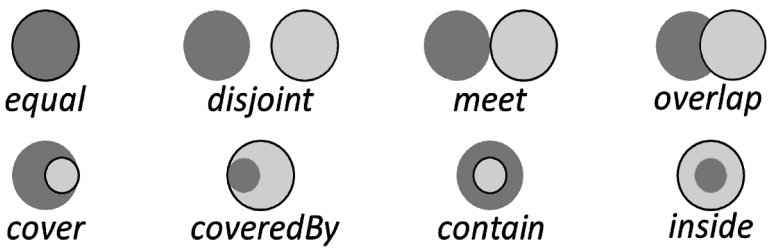


Table 1. 9-intersection and extracted geographical relationships

9-Intersection	Extracted Geographical Relationships
equal	*equal*
disjoint	*disjoint*
meet	*meet*
overlap	*overlap*
cover	*contain*
contain	
coveredBy	*inside*
inside	

Figure 5 shows an example of scene detection when γ is set to 0.8. The columns of the table denote extracted topics and their *tf-idf* values are in descending order from the left to the right. The rows denote sections of closed captions in the ascending order of time sequence from the top to the bottom. In this example, scenes are detected as two frames since their total *tf-idf* values of extracted topics along the timeline of the video are higher than 0.8.

Measuring Popularity Rating of Scenes

In order to measure popularity rating of scenes in a video, the authors first use the topics K_j of each scene j as a query to acquire the number of search hits of each scene $|Search(K_j)|$ from online video sharing sites such as YouTube. Next, the authors calculate a threshold value δ to measure the popularity rating of each scene as follows:

$$\begin{cases} |Search(K_j)| \geq \delta, high\ popularity\ rating \\ |Search(K_j)| < \delta,\ low\ popularity\ rating \end{cases} \tag{5}$$

$$\delta = \frac{|Search(K)|}{N} \tag{6}$$

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

MEDIA SYNCHRONIZATION WITH TV PROGRAMS

Extracting Semantic Structure

In order to present geographical contents with TV programs, the authors extract semantic structure, which is the intentions of scenes in a video, based on geographical metadata of the video shown in Table 2. This table shows the intentions determined by temporal sequences and geographical relationships between two given locations, *A* and *B*, appeared in the video.

The semantic structure is extracted as *description of a single spot* when the relationship that *A equals A* appears in one scene. The authors assume that the intention of this scene is to describe *A*. The semantic structure is extracted as *description of spots* when the relationship that *A contains B* appears in one scene. The authors assume that the intention of this scene is to describe both *A* and *B*, and *A* includes *B*. The semantic structure is extracted as *comparison of spots* when the relationship that *A*

Figure 5. An example of scene detection based on topics of a video

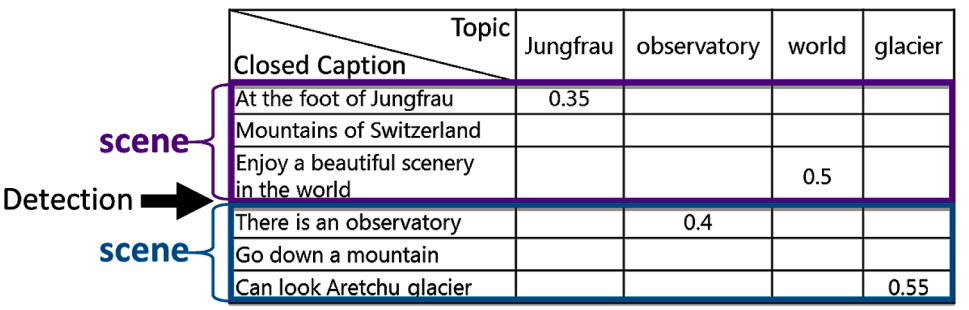


Table 2. Semantic structure based on geographical metadata

Intentions	Temporal Sequences	Geographical Relationships
Description of a single spot	Location *A* → Location *A*	*equal*
Description of spots	Location *A* → Location *B*	*contain*
Comparison of spots		*disjoint* (*in wide regions*)
		meet
Description of a route		*disjoint* (*in narrow regions*)
Description of regions		*overlap*
		inside

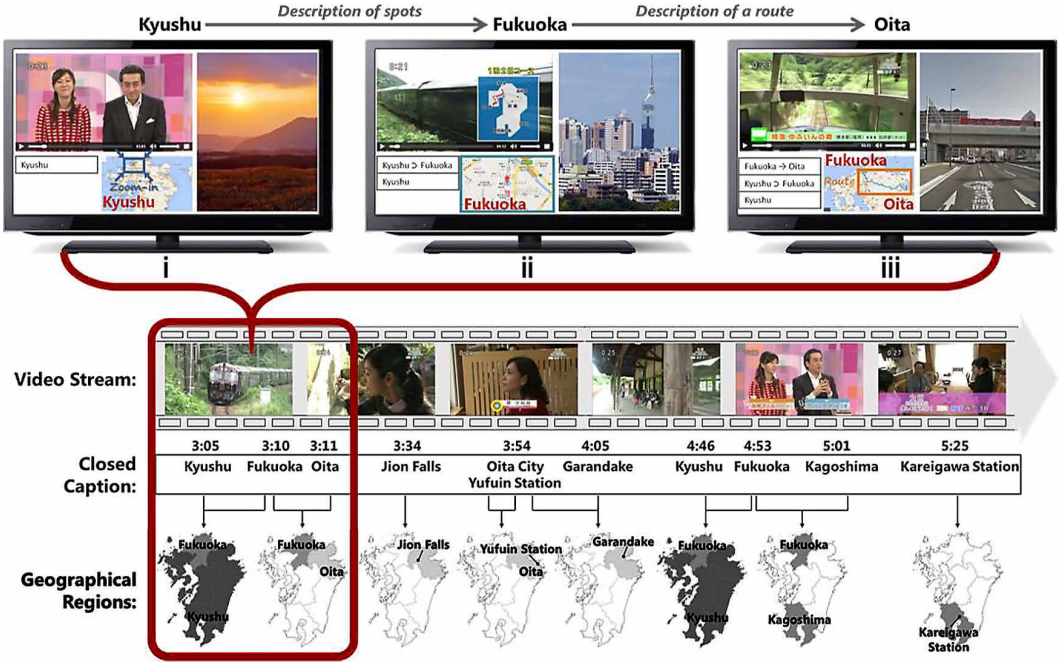
and *B* are *disjoint* appears in one scene and each location has its own wide region in the same level (e.g., prefectures), or the relationship that *A meets B* appears in one scene. The authors assume that the intention of this scene is to compare *A* and *B*. The semantic structure is extracted as *description of a route* when the relationship that *A* and *B* are *disjoint* appears in one scene and each location has its own narrow region in the same level (e.g., municipalities). The authors assume that the intention of this scene is to describe a route from *A* to *B*. The semantic structure is extracted as *description of regions* when the relationship that *A overlaps B* or *A* is *inside B* appears in one scene. The authors assume that the intention of this scene is to describe the overlapping area of *A* and *B*.

Presenting Geographical Contents Synchronized with TV Programs

When the intention of a video is to describe spots, the media synchronization mechanism presents maps of given locations and related photos seamlessly in order to help viewers easily grasp geographical positions and appearances of these locations. When the intention of a video is to compare spots, the media synchronization mechanism presents a map including all given locations and their photos seamlessly in order to help viewers better understand positional relationships and their appearances. When the intention of a video is to describe routes, the media synchronization mechanism presents a map and Street View of the route between given locations seamlessly in order to help viewers easily know the route and the distance between them. When the intention of a video is to describe regions, the media synchronization mechanism presents the map of given locations and the photos of the overlapping region of the given locations seamlessly in order to help viewers easily grasp the locations and the atmosphere of their common region.

For example, suppose that three location names, Kyushu (2:59), Fukuoka (3:11), Oita (3:17), will appear in a video (see Figure 6). When the first two location names are detected in one scene, the media synchronization mechanism determines their intention and presents corresponding geographic contents synchronized with the video. Specifically, the system extracts the two geographical relationships: Kyushu *contains* Fukuoka, Fukuoka and Oita are *disjoint*. In Figure 6, when the video shows Kyushu first and then Fukuoka in the same scene at the first time in a video player window, a map window presents the region of Kyushu first and then zooms in to the region of Fukuoka. Meanwhile, a Street View window presents photos of Kyushu as shown in **i**. Then, when the video shows Fukuoka, the Street View window presents photos of Fukuoka as shown in **ii**. When the video shows Oita after Fukuoka in the same scene, the map window presents both regions of Fukuoka and Oita, and highlights a route between them. At the same time, the Street View starts at Fukuoka and proceeds along the route to Oita as shown in **iii**. In this manner, the media synchronization mechanism enables viewers to easily and effectively grasp the geographical information of locations appeared in closed captions of videos.

Figure 6. An example of presenting geographic contents synchronized with a video



VIDEO RECONSTRUCTION FROM TV PROGRAMS

Determining Additional and Removal Contents

In order to generate new video contents from a TV program, the authors determine which original scenes should be added with other related contents, and which original scenes should be removed. In this work, for experts the authors add other related contents after original scenes with *low* popularity ratings and remove original scenes with *high* popularity ratings to generate a detailed video or a digest video (**P1**, **P3**). On the contrary, for non-experts the authors add other related contents after original scenes with *high* popularity ratings and remove original scenes with *low* popularity ratings to generate a detailed video or a digest video (**P2**, **P4**). In other words, new video contents can be generated based on the popularity rating of each original scene with the following conditions:

- **P1** and **P3** (for experts):
  - If *a scene is with a low popularity rating* then *add other related contents*;
  - Else if *the scene is with a high popularity rating* then *remove this scene*.
- **P2** and **P4** (for non-experts):
  - If *a scene is with a high popularity rating* then *add other related contents*;
  - Else if *the scene is with a low popularity rating* then *remove this scene*.

Here, original scenes keep as they are in the video if they do not satisfy the above conditions.

Furthermore, the system selects additional contents based on popularity ratings of original scenes as follows. If the additional contents are YouTube videos, the system muffles their voice so that the generated video contents can be integrated into the original video in a smooth way:

- **P1** and **P3** (for experts): YouTube videos, Google Images, or detailed maps from Google Maps based on relevance ranking by searching topics of original scenes with low popularity ratings;
- **P2** and **P4** (for non-experts): YouTube videos, Google Images, or extensive maps from Google Maps based on relevance ranking by searching topics of original scenes with high popularity ratings.

Calculating Viewing Time of Additional Contents

In order to calculate the viewing time t_j of additional contents and control LOD of generated video contents, the system calculates the ratio of the popularity rating of each original scene $|Search(K_j)|$ to the total number of original scenes with low or high popularity ratings denoted as N_l or N_h:

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

$$t_j = \frac{\left|Search\left(K_j\right)\right|}{N_l} \times T \tag{7}$$

Here, N_l denotes the total number of original scenes whose popularity ratings are lower than δ defined in Equation (6). T is the total viewing time of additional contents:

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

$$t_j = \frac{\left|Search\left(K_j\right)\right|}{N_h} \times T \tag{8}$$

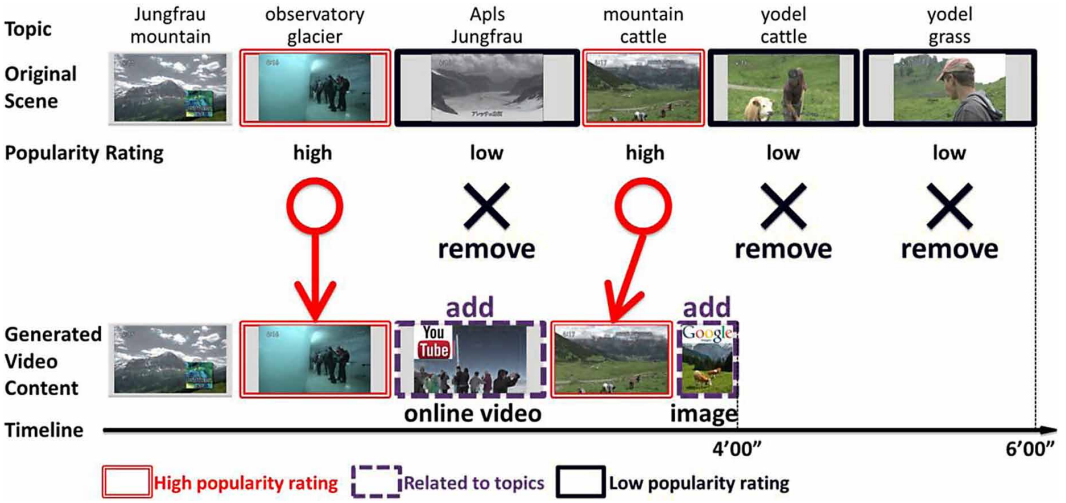
Here, N_h denotes the total number of original scenes whose popularity ratings are higher exceed δ defined in Equation (6).

In addition, the system determines the types of additional contents according to their viewing time by using a threshold value θ based on time pressure. θ is set to the shortest time of sections of closed captions in an original video:

- If $t_j \geq \theta$ then *add YouTube videos or Google Maps* (*if topics are location names*);
- If $t_j < \theta$ then *add Google Images or Google Maps* (*if topics are location names*).

An example is shown in Figure 7. This figure depicts an overview of generating digest videos for non-expert viewers (**P4**) by the video reconstruction mechanism. Scenes with high or low popularity ratings are detected by extracting topics from closed captions of a video. Double line frames denote original scenes with high popularity ratings. Dashed line frames denote additional contents related to topics of original scenes with high popularity ratings. Single line frames denote original scenes with low popularity ratings, which will be removed from the original video. On the other hand, the leftmost scene will not be modified. In this manner, if scenes of a video are with low popularity ratings and viewers want to gain more information about their interested topics, they can watch a full-length detailed video for experts (**P1**); if scenes of a video are with high popularity ratings and viewers want to grasp a summary of their interested topics, they can watch a digest video for non-expert viewers (**P4**).

Figure 7. An example of generating P4 from a video



USER STUDY AND DISCUSSION

Evaluating Presentation of Geographical Contents with TV Programs

The purpose of this user study is to evaluate the effectiveness of presenting geographic contents. The authors evaluate three patterns of geographic contents presentation using the proposed media synchronization mechanism, including geographical relationships between locations, routes between locations, and geographical positions of single locations. The video used for evaluation is from a travel program (Kyushu, 2013) introducing Kyushu in Japan with a total viewing time of 23 minutes. The example in Figure 6 is a part of this video. Location names appeared in the video include Kyushu (2:59), Fukuoka (3:11), Oita (3:17), and Jion Falls (3:34).

The study is completed by five college students, who have never been to Kyushu. They evaluate the effectiveness of presenting geographic contents based on a five-level Likert scale (1: very ineffective ~ 3: neither effective nor ineffective ~ 5: very effective). The results and findings are summarized as follows:

- The first pattern is the description of spots on geographical relationships between them. When the video shows Kyushu first and then Fukuoka in a 12 seconds scene, the system presents a map of the region of Kyushu first and then zooms in the map to the region of Fukuoka. The average result of this pattern is 4.8. This demonstrates that it can greatly help subjects understand geographical positions of Kyushu and Fukuoka, and their geographical relationship;
- The second pattern is the description of a route between two locations. When the video shows Fukuoka first and then Oita in a 6 seconds scene, the system presents a map of both Fukuoka and Oita, and shows a route between them. The average result of this pattern is 4.4, which means that it is good to help subjects grasp geographical positions of Fukuoka and Oita, and the route between them;
- The third pattern is the description of a single spot, i.e., a location. When the video shows Jion Falls continuously in a 17 seconds scene, the system zooms in the map of Fukuoka and Oita in the previous scene to the region of Jion Falls. The average result of this pattern is 4.0. This shows that it can help subjects understand the geographical position of Jion Falls. In this pattern, the system also presents online photos of Jion Falls. The average result is 3.2, which means that it is a little hard to rouse the subjects' interests by presenting online photos.

As discussed above, the presented online photos are not very effective. The authors consider improving it by using a 3D model to describe the locations by adopting Google Earth.

Evaluating Generation of Video Contents from TV Programs

The purpose of this evaluation with two experiments is to verify whether the proposed video reconstruction mechanism is useful for helping viewers to enjoy videos in TV programs.

Experimental Dataset

The authors used three videos from NHK World Heritage 100 (World Heritage, 2014) and generated four types of videos **P1** ~ **P4**, respectively:

- **Original video (V1):** Swiss Alps Jungfrau-Aletsch, Switzerland (viewing time: 5'31");
- **Original video (V2):** Wachau Cultural Landscape, Austria (viewing time: 5'30");
- **Original video (V3):** Yellowstone National Park, United States (viewing time: 5'30").

Table 3 shows the generated video contents in Experiment I by using only videos as additional contents. Table 4 shows the generated video contents in Experiment II by using videos, images or maps as additional contents. Here, '+' denotes 'added' and '-' denotes 'removed'. Furthermore, the authors determine the types of additional contents according to their viewing time using a threshold value of $\theta = 10$. The number of scenes detected by the system is 9 for V1, 11 for V2, and 13 for V3, by using a threshold value of $\gamma = 0.8$. The average viewing time of detected scenes is approximately 30 seconds. In addition, the average viewing time of detailed videos (**P1**, **P2**) is 8'22", and that of digest videos (**P3**, **P4**) is 2'32" in the two experiments.

Ten college students participated in the experiments. They completed the following 6 items (**Content Understanding:** *Q1*, **Editing Effects:** *Q2* ~ *Q4*, **Interest-Arousing:** *Q5*, *Q6*) in a

Table 3. 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 4. 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)

questionnaire after they watched the four types of videos P1 ~ P4 generated from V1, V2, and V3 in experiments I and II, respectively:

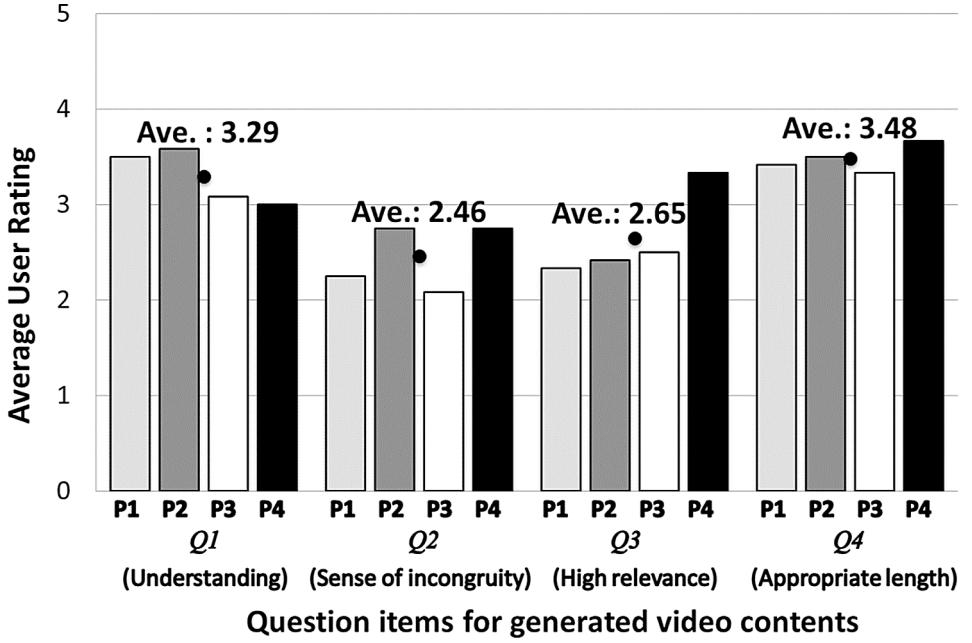
Q1: Could understand the video contents.
Q2: No sense of incongruity in switching scenes.
Q3: Unrelated scenes do not exist.
Q4: Felt long about the video content of **P1** or **P2**. Felt short about the video content of **P3** or **P4**.
Q5: Write down the topics that you are interested in and their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

Experiment I: Generation of Video Contents by Adding Videos

Figure 8 illustrates the average rating of *Q1* ~ *Q4* in Experiment I based on a five-level Likert scale. High user ratings indicate good results. The results and findings are shown as follows:

- *Q1* for detailed videos (**P1**, **P2**) gain high ratings, while *Q1* for digest videos (**P3**, **P4**) are rated low because several important scenes are removed;
- *Q2* for **P1** ~ **P4** are rated low. Although the original videos contain narration, the added video contents do not contain narration. This caused subjects to feel a sense of incongruity in switching scenes. In particular, the generated video contents for experts (**P1**, **P3**) are rated low;
- *Q3* for digest videos for non-expert viewers (**P4**) reach high ratings. However, detailed videos (**P1**, **P2**) are rated low, since unrelated scenes are added;
- In *Q4* for detailed videos (**P1**, **P2**), many subjects felt long. *Q4* for digest videos for non-expert viewers (**P4**) obtain the highest ratings because most of subjects felt short;
- In *Q5* for the interested topics written by subjects, 82% of these topics are detected by the proposed video reconstruction mechanism. Moreover, there are several topics that subjects are not interested in when they watch the original videos, but become interested in after watching the

Figure 8. Experiment I: The results of Q1 ~ Q4 in questionnaire



generated video contents. Thus, the proposed video reconstruction mechanism is able to arouse the subjects' interests, and can help subjects enjoy the videos;

- In *Q6*, 24 topics are reported to be not related to detailed videos (**P1**, **P2**) and 18 topics are reported to be not related to digest videos (**P3**, **P4**). Meanwhile, even though some topics are related to the original videos, but the added scenes are not appropriate.

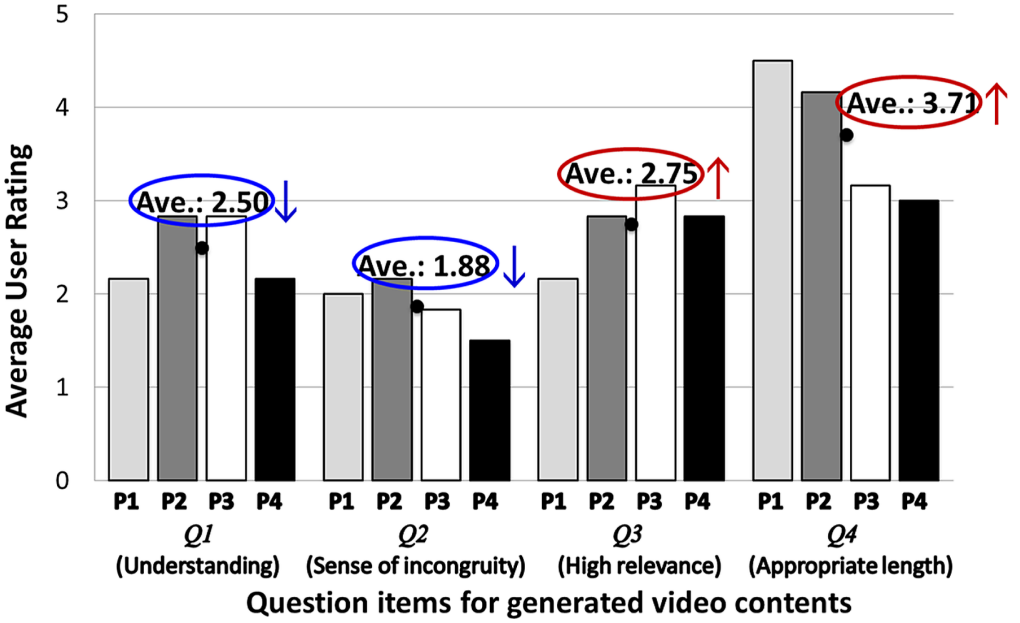
As discussed above, this experiment shows that *Q1* and *Q4* for generated video contents gain high ratings. *Q2* for generated video contents are rated low. In particular, *Q3* for **P1**, **P2**, and **P3** are rated low.

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

Figure 9 illustrates the average rating of *Q1 ~ Q4* in Experiment II. The results are compared with Experiment I and the findings are summarized as follows:

- *Q1* for **P1** and **P4** are rated low because several important scenes are removed as in Experiment I;
- *Q2* for **P1 ~ P4** are rated low and the average rating of *Q2* is lower than that of Experiment I. That is because static images such as photos or maps are increased, and subjects strongly felt a sense of incongruity when they watch the static images during the video;
- *Q3* for detailed videos for experts (**P1**) are also rated low, but the average rating of *Q3* is higher than that of Experiment I. The reason is that related contents became more detailed by using various types of additional contents;
- *Q4* for detailed videos (**P1**, **P2**) gain high ratings because most of subjects felt long. The similar is to digest videos (**P3**, **P4**) because many subjects felt short. Therefore, the authors confirm that the viewing time of generated video contents performs as well as in Experiment I;

Figure 9. Experiment II: The results of Q1 ~ Q4 in questionnaire



- In *Q5* for the interested topics written by subjects, 78% of these topics are detected by the proposed video reconstruction mechanism. Thus, the proposed video reconstruction mechanism is able to arouse the subjects' interests as in Experiment I;
- In *Q6*, 12 topics are reported to be not related to detailed videos (**P1**, **P2**) and 8 topics are reported to be not related to digest videos (**P3**, **P4**). Compared with Experiment I, there are fewer topics that are reported to be not related to the generated videos.

In summary, this experiment shows that the average ratings of *Q1* and *Q2* for **P1** ~ **P4** in Experiment II are lower than those of Experiment I. *Q3* and *Q6* for **P1** ~ **P4** in Experiment II show good results. The results of *Q4* and *Q5* are the similar between Experiments I and II. In the future, it is necessary to consider how to switch scenes smoothly. Furthermore, the authors need to analyze the relevance between additional contents and original videos to keep the main story of the original video smooth, and check the validity of removed original scenes for generating digest videos.

CONCLUSION AND FUTURE WORK

In this paper, the authors developed a novel automatic video reinforcing system with two functions: (1) a media synchronization mechanism, which automatically presents geographic contents synchronized with a video based on geographical metadata and semantic structure of the video; (2) a video reconstruction mechanism, which automatically generates four kinds of video contents from a video by adding other contents and removing original scenes, based on topical metadata and LOD controlling under time pressure. In order to extract geographical metadata and topical metadata of videos, the authors utilize closed captions of the videos. The system extracts geographical relationships of location names appeared in each scene, and measures popularity ratings of scenes by calculating the number of search hits of topics appeared in each scene. Finally, the authors conducted (1) a user study with the proposed media synchronization mechanism, and (2) two experiments with the proposed video reconstruction mechanism. The results of user study showed that the proposed

media synchronization mechanism can help viewers effectively grasp geographic information of videos from travel programs through simultaneously viewing geographic contents with the videos. The experimental results revealed that the proposed video reconstruction mechanism can help users enjoy the video contents that suit viewers' interests.

In the future, the authors plan to improve the methods for presenting geographic contents by considering cinematography and film languages, and consider selecting additional contents of other types (e.g., voice, web pages, and microblogs). Another future direction is to extend the system by user interactions. For example, the system can be extended to allow viewers to decide whether and how to show the supplementary information and allow them to control the viewing time.

ACKNOWLEDGMENT

This work was partially supported by SCOPE of the Ministry of Internal Affairs and Communications of Japan, and JSPS KAKENHI Grant Numbers 26280042, 15K00162, 25280039.

REFERENCES

Banjou, Y., Takakura, H., & Kambayashi, Y. (1997). Generation of rough maps according to various user requirements. *Proceedings of the 55th National Convention of Information Processing Society of Japan IPSJ 55* (pp. 481-482). IPSJ, Fukuoka, Japan. (in Japanese)

Baraldi, L., Grana, C., & Cucchiara, R. (2015). Scene segmentation using temporal clustering for accessing and re-using broadcast video. *Proceedings of the 2015 IEEE International Conference on Multimedia and Expo (ICME 2015)*. Torino, Italy: IEEE. doi:10.1109/ICME.2015.7177476

Chakraborty, S., Tickoo, O., & Iyer, R. (2014). Adaptive keyframe selection for video summarization. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV 2015)*, Hawaii, USA. IEEE.

Chi, T. H., Lo, H. H., Chu, Y. H., & Lin, W. C. (2009). A mobile tourism application model based on collective interactive genetic algorithms. *Proceedings of the 12th International Conference on Computer and Information Technology (ICCIT 2009)*, Dhaka, Bangladesh. IEEE. doi:10.1109/ICCIT.2009.280

Evers, M., Theune, M., & Karreman, J. (2007). Which way to turn?: Guide orientation in virtual way finding. *Proceedings of the Workshop on Embodied Language Processing (EmbodiedNLP 2007)*. Prague, Czech Republic: Association for Computational Linguistics. doi:10.3115/1610065.1610069

Fabro, D. M., Schoeffmann, K., & Böszörmenyi, L. (2010). Instant video browsing: A tool for fast non-sequential hierarchical video browsing. *HCI in Work and Learning*. *Life and Leisure*, *53*(2), 391–429.

Fleites, F., Wang, H., & Chen, S.-C. (2015a). Enhancing product detection with multicue optimization for TV shopping applications. *Transactions on Emerging Topics in Computing*, *3*(2), 161–171. doi:10.1109/TETC.2014.2386140

Fleites, F., Wang, H., & Chen, S.-C. (2015b). Enabling enriched TV shopping experience via computational and temporal aware view-centric multimedia abstraction. *IEEE Transactions on Multimedia*, *17*(7), 1068–1080. doi:10.1109/TMM.2015.2433213

Foulke, E., & Sticht, T. G. (1969). Review of research on the intelligibility and comprehension of accelerated speech. *Speech and Hearing Science: Selected Readings*, *72*(1), 50–62. PMID:4897155

Geerts, D., Leenheer, R. D., Grooff, D., Negenman, J., & Heijstraten, S. (2014). In front of and behind the second screen: Viewer and producer perspectives on a companion app. *Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video* (*TVX 2014*). Newcastle Upon Tyne, UK: ACM.

Google Maps. (2016). Google maps API. *Google Developers*. Retrieved from https://developers.google.com/maps/

Google Street View Imagery. (2016). Google street view image API. *Google Developers*. Retrieved from https://developers.google.com/maps/documentation/streetview/

Haklay, M., & Martin, H. (2008). OpenStreetMap: User-generated street maps. *IEEE Pervasive Computing / IEEE Computer Society [and] IEEE Communications Society*, *7*(4), 12–18. doi:10.1109/MPRV.2008.80

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

Kawamura, S., Fukusato, T., Hirai, T., & Morishima, S. (2014). Efficient video viewing system for racquet sports with automatic summarization focusing on rally scenes. *Proceedings of the 41st International Conference and Exhibition on Computer Graphics and Interactive Techniques* (*SIGGRAPH 2014*). Vancouver, Canada: ACM. doi:10.1145/2614217.2614240

Kitayama, D., Miyamoto, S., & Sumiya, K. (2010). A customizing method of digital maps based on user's operations and object's appearing patterns. *Transactions of Information Processing Society of Japan, 3*(4), 65-81. (in Japanese)

Kobayashi, K., Kitayama, D., & Sumiya, K. (2011). Cinematic street: Automatic street view walk-through system using modified maps. *Proceedings of the 10th International Symposium on Web & Geographical Information Systems* (*W2GIS 2011*), Kyoto, Japan. Springer. doi:10.1007/978-3-642-19173-2_12

Kurata, Y. (2010). An overview of the research on topological relations and future issues in GIScience [in Japanese]. *Theory and Applications of GIS*, *18*(2), 41–51.

Kurihara, K. (2012). CinemaGazer: a system for watching videos at very high speed. *Proceedings of the International Working Conference on Advanced Visual Interfaces* (A*VI 2012*). Naples, Italy: ACM. doi:10.1145/2254556.2254579

Kyushu. (2013). Kyushu super luxury train go the birth beauty of the country. *NHK*. Retrieved from http://www.nhk.or.jp/bs-blog/100/172311.html

Liu, C., Huang, Q., Jiang, S., Xing, L., Ye, Q., & Gao, W. (2009). A framework for flexible summarization of racquet sports video using multiple modalities. *Computer Vision and Image Understanding*, *113*(3), 415–424. doi:10.1016/j.cviu.2008.08.002

Liu, C., Wang, D., Zhu, J., & Zhang, B. (2013). Learning a contextual multi-thread model for movie/TV scene segmentation. *IEEE Transactions on Multimedia*, *15*(4), 884–897. doi:10.1109/TMM.2013.2238522

Nandakumar, A., & Murray, J. (2014). Companion apps for long arc TV series: Supporting new viewers in complex storyworlds with tightly synchronized context-sensitive annotations. *Proceedings of the 2014 ACM international conference on Interactive experiences for TV and online video* (*TVX 2014*), Newcastle Upon Tyne, UK. ACM. doi:10.1145/2602299.2602317

Panoramio (2016). Panoramio API. *Google Developers*. Retrieved from http://www.panoramio.com/api/data/api.html

Rasheed, Z., & Shah, M. (2005). Detection and representation of scenes in videos. *IEEE Transactions on Multimedia*, *7*(6), 1097–1105. doi:10.1109/TMM.2005.858392

Song, Y., Ogawa, T., & Haseyama, M. (2014). A scene segmentation approach based on the mcmc method using video structures. [in Japanese]. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Science*, *97*(3), 560–573.

Takahashi, A., Tatedoko, M., Shimizu, T., Kinutani, H., & Yoshikawa, M. (2010). Metadata management for integration and analysis of earth observation data. *Journal of Software*, *5*(2), 168–178.

Viana, W., Miron, A. D., Moisuc, B., Gensel, J., Villanova-Oliver, M., & Martin, H. (2011). Towards the semantic and context-aware management of mobile multimedia. *Multimedia Tools and Applications*, *53*(2), 391–429. doi:10.1007/s11042-010-0502-6

Wang, Y., Kawai, Y., Sumiya, K., & Ishikawa, Y. (2015). An automatic video reinforcing system based on popularity rating of scenes and level of detail controlling. *Proceedings of the 2015 IEEE International Symposium on Multimedia* (*ISM 2015*), Miami, FL, USA. IEEE.

World Heritage. (2014). Series of world heritage 100. *NHK*. Retrieved from http://www4.nhk.or.jp/P166/

Yeung, M., Yeo, B.-L., & Liu, B. (1998). Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding*, *71*(1), 94-109. doi:10.1006/cviu.1997.0628

Yuanyuan Wang is an assistant professor at Graduate School of Science and Engineering, Yamaguchi University, Japan. She received her MS and PhD degrees from Human Science and Environment from University of Hyogo in 2011 and 2014, respectively. She has worked as a research intern at the Human-Computer Interaction Group, Microsoft Research Asia in 2013, a researcher at Faculty of Computer Science and Engineering, Kyoto Sangyo University in 2014, and a research associate at Graduate School of Information Science, Nagoya University in 2015. Her research interests include e-learning systems, structural analysis, multimedia databases, and human-computer interaction. She is a member of Information Processing Society of Japan (IPSJ) and the Database Society of Japan (DBSJ).

Daisuke Kitayama is an assistant professor at Faculty of Information Studies, Kogakuin University, Japan. His research interests include geographical database and user interaction. He received his MS and PhD degrees from Human Science and Environment from University of Hyogo in 2007 and 2009, respectively. He is a member of Association for Computing Machinery (ACM), the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), and the Database Society of Japan (DBSJ).

Yukiko Kawai received the MS and PhD degrees in Information Science and Technology from Nara Institute of Science and Technology, in 1999 and 2001, respectively. She has worked as a research fellow at the National Institute of Information and Communications Technology from 2001 to 2006. She has been a lecturer at Kyoto Sangyo University from 2006 to 2011, and she is currently an associate professor at Kyoto Sangyo University since 2011. Her research interests include data mining, information analyzing and Web information retrieval. She is a member of the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), and the Database Society of Japan (DBSJ).

Kazutoshi Sumiya is a professor at School of Policy Studies, Kwansei Gakuin University. His research interests include multimedia databases, Web services, information retrieval, and geographical information systems. He received his PhD in information media from Kobe University and worked for Panasonic in Japan. He was an associate professor at School of Informatics, Kyoto University, and a professor at School of Human Science and Environment, University of Hyogo. He is a member of IEEE Computer Society, Association for Computing Machinery (ACM), Information Processing Society of Japan (IPSJ), Database Society of Japan (DBSJ) and Institute of Electronics, Information and Communication Engineers (IEICE).

Yoshiharu Ishikawa is a professor in Graduate School of Information Science, Nagoya University. He received BS, ME, and DrEng Degrees from University of Tsukuba in 1989, 1991, and 1995, respectively. His research interests include spatio-temporal databases, data mining, information retrieval, and Web information systems. He is a member of Association for Computing Machinery (ACM), IEEE Computer Society, the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), the Database Society of Japan and the Institute of Electronics (DBSJ), and The Japanese Society for Artificial Intelligence (JSAI).