An Automatic Media Synchronizing Mechanism with TV Programs

Yuanyuan Wang
Nagoya University
Furo-cho
Chikusa-ku, Nagoya
464-8601 Japan
yuanw@dd.ss.is.nagoya-u.ac.jp

Kazutoshi Sumiya
Kwansei Gakunin University
2-1 Gakuen
Sanda, Hyogo
669-1337 Japan
sumiya@kwansei.ac.jp

Yukiko Kawai
Kyoto Sangyo University
Motoyama, Kamigamo
Kita-ku, Kyoto
603-8555 Japan
kawai@cc.kyoto-su.ac.jp

MediaSync 2015 in conjunction with ACM TVX 2015, June 3, 2015, Brussels, Belgium.

Abstract

Almost TV programs provide a high affinity geographic data, such as tourist spots, historical places. However, current services cannot present geographic contents synchronized with the TV programs, and viewers (e.g., tourists) difficult to grasp the surroundings of the geographic data, how the locations are related, and distances between them during the TV programs. Therefore, we have developed a system based on the concept of a second screen service, which presents supplementary information synchronized with a TV program on a big display or smartphones, this enables the users to easily understand the geographic data of the TV programs. When one scene first introduces New York, next, touring Manhattan; the system simultaneously presents a map of both, and moving on the route between them. For this, the system first extracts geographical geometry, i.e., locations, geographical relationships, and semantic structure, i.e., temporal meaning, intentions, by extracting place names and their appearance time from closed captions of video streams. Based on them, the system presents geographic contents (i.e., maps, Street View, etc.) with seamless effects.

Author Keywords

TV program; geographical geometry; semantic structure; closed captions

Introduction

As well as a new second screen service, NHK's Hybridcast [1] enables TV programs to be integrated with Web content, e.g., news, weather. Viewers can check current weather of tourist spots by manual during a travel program. Various TV programs, such as travel program and educational program, mainly provide a high affinity geographic data, such as tourist spots or historical places as topics in video streams. However, current services cannot present geographic contents synchronized with the TV programs, and viewers (e.g., tourists, children, etc.) difficult to grasp the surroundings of the geographic data, how the locations are related, and distances between them during the TV programs. For instance, a travel program introduces tourist spots in Kyoto, Japan. When one scene shows a restaurant in Gion, the next scene is touring Yasaka, the viewers were unobserved the moving between Gion and Yasaka, and they cannot directly understand where and how about distances between them during the travel program.

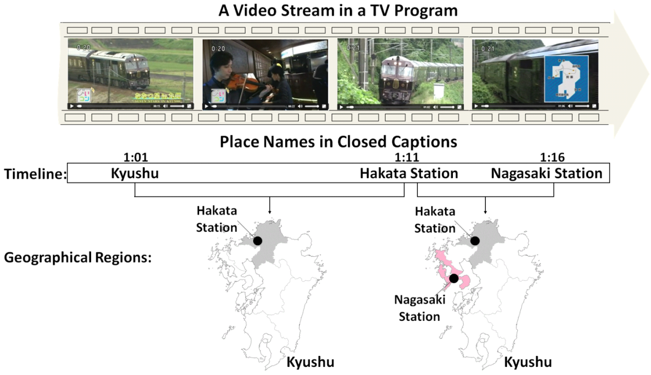

Figure 1: Geographical geometry of a video stream

The growing success of TV programs using tablets and smartphones as second screen devices [2], and the second screen service often offers extra information about the TV programs [6]. Therefore, we aim to develop a system that provides a TeleVision Map Interface for Location AwareNess, called TV-Milan, based on the concept of the second screen service. For this, the system automatically extracts geographical geometry and semantic structure of video streams. Figure 1 shows our core research model of the geographical geometry, it can be implemented by 1) extracting the temporal sequences of place names which appear in closed captions of the video stream; and 2) extracting geographical relationships between the places based on their geographical regions. On the basis of the geographical geometry, we determine semantic structure such as the intentions of the video stream, to determine how to present geographic contents (i.e., maps, Street View, etc.) with seamless effects during the video stream. Therefore, the TV-Milan presents geographic contents or Web contents (i.e., photos, Web pages, video clips, etc.) synchronized with the TV programs on a big display (see Figure 2) or smartphones, which aids users to grasp geographical information (i.e., tourist spots, routes, etc.) in detail during the video stream easily and efficiently.

With our TV-Milan, a route between two places can be synchronously moved with the video stream by using Street View[2], when names of these two places appear in a closed caption of one scene. In addition, in order to grasp the surroundings of the places and how about the locations, distances between them based on the semantic structure of the video stream, the TV-Milan does not present Street View only, but also Google Map or Google Earth[3] with Web contents, such as online photos, video

[1] http://www.nhk.or.jp/hybridcast/online/

[2] https://www.google.com/maps/views/?gl=jp

[3] http://earth.google.co.jp/

clips, and so on. Therefore, the TV-Milan can enable users easily and efficiently grasp the geographical information of the places, which appeared but not described in the video streams.

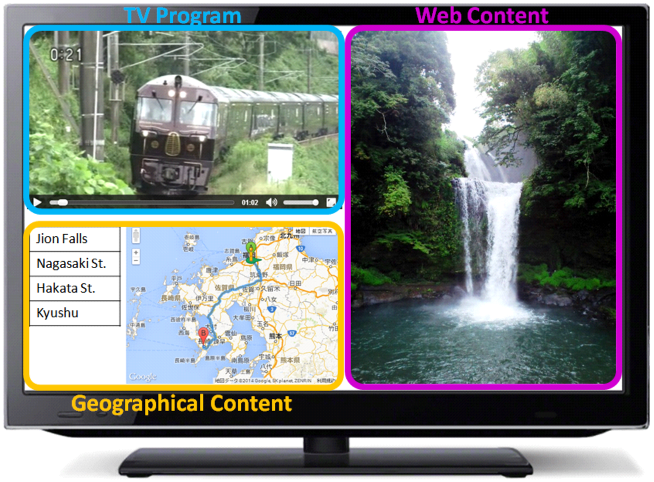

Figure 2: Conceptual diagram of TV-Milan

The next section describes a preliminary survey to deepen our understanding of presentation of geographic contents and reviews related work. Section 3 explains our research model by analyzing video streams. Section 4 presents a media synchronizing mechanism with TV Programs. Finally, Section 6 concludes this paper with future works.

Preliminary Study and Related Work

Preliminary Study

We conducted a simple user study to learn about presentation of geographic contents with seamless effects for understanding the semantic structure that the intentions of a video stream. We prepared three patterns of geographic content presentation from a real video stream in a play order by using our developed prototype system, geographical relationships between places, a route between places, and a location of a single place. Place names, i.e., Kyushu (3:05), Fukuoka (3:10), Oita (3:11), and so on, they appear in the video stream, which is a travel program[4] introducing Kyushu in Japan, with total run-time of 23 minutes.

There were five college students (all males) who never been Kyushu and who were not interested in travel, completed the survey about the effectiveness of the presented geographic contents with seamless effects by using a five-level Likert scale (1: very ineffective∼3: neither effective nor ineffective∼5: very effective); the results and our findings were summarized as follows:

- The pattern one was introduction to places about geographical relationships between them, when the video shows Kyushu to Fukuoka in a 15 second scene, simultaneously, the prototype system presents a map of the region of Kyushu and zooms-in to the region of Fukuoka during this 15 second scene. The average point of this pattern was 4.8, it was great to help the participants understand the locations of Kyushu and Fukuoka, and the geographical relationships of them.
- The pattern two was introduction to a route, when the video shows Fukuoka to Oita in a 15 second scene, simultaneously, the prototype system presents a map of both Fukuoka and Oita, and moves on the route between them during this 15 second scene. The average point of this pattern was 4.4, it was good to help the participants grasp the locations of Fukuoka and Oita, and the route between them.

[4] http://www.nhk.or.jp/hakken/kyushu/program/tetsudo.html

- The pattern three was introduction to a single place about a location of it, when the video shows Jion Falls continuously in a 15 second scene, simultaneously, the prototype system presents a map of both Fukuoka and Oita in the previous scene and zooms-in to the region of Jion Falls during this 15 second scene. The average point of this pattern was 4.0, it was good to help the participants understand the location of Jion Falls. In this pattern, the prototype system also presents the photos related to Jion Falls. The average point of evaluation results was 3.2, it was a little hard to rouse the users' interests.

Related Work

Many studies extract and utilize geographical information from various contents. Takahashi et al. [8] proposed a ranking method applied to earth science data. They extracted temporal information and spatial information from articles based on the link structure of Wikipedia. Kitayama et al. [3] 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 place names. Our TV-Milan is similar to these works; we utilize geographical information of TV programs, to present geographic contents supplement for understanding the TV programs.

Some studies have tried to provide location-aware interface by exploring the geographical information to help users in terms of localization. Viana et al. [9] proposed using context awareness and semantic technologies in order to improve and facilitate the organization, annotation, retrieval and sharing of personal mobile multimedia documents. This approach combines metadata that are extracted and enriched automatically from the users' context, i.e., locations, spatial relationships, etc.. They are similar to our work that we propose an interface for location awareness by extracting geographical geometry such as locations, geographical relationships.

As in related works about generation of various contents according to the purposes of users. Banjou et al. [1] proposed a method for generating rough map according to the purposes by interactive trial and error with the users' requirements. Kobayashi et al. [4] proposed a system for transforming a modified map into a streaming video based on the intentions of the map's maker to express what he wants to present. In this work, we propose a television map interface for presenting geographic contents synchronized with the TV program, which utilizes geographical information in video streams by considering the semantic structure of the video streams.

Geographical Geometry and Semantic Structure of Video Streams

In this work, we extract geographical geometry of video streams, including temporal sequences and geographical relationships, by detecting place names from video streams. Based on them, we then extract semantic structure that the intentions of the video streams.

There are two methods to detect place names from a video stream, one is extracting the place names in closed captions by analyzing a MPEG-2 Transport Stream file (with the HbbTV standard) of the video stream, and the other one is identifying the place names of objects in video scenes by using an image-recognition technique. In this work, our goal is to supplement the video stream that contains the place names but does not describe them.

Therefore, we extract the temporal sequence that an order of appearance of the place names from the closed caption.

We consider that one video stream, often consist of various topics, then, we need to classify scenes according to the topic by focusing geographic data and a time width of each scene. For this, we change all place names of the video stream into latitude and longitude coordinates, and detect latitude and longitude coordinates (X, Y) of a center point c of all places by the following formula:

$$(X, Y) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n} \right)$$

Here, n place names appeared in the video stream, and i denotes one place that the number of this place in the video stream. x_i denotes a latitude coordinate of i, and y_i denotes a longitude coordinate of i.

Therefore, we determine one scene of the video stream by considering both the geographical distance and temporal distance between every two places in their appearance order. In addition, we consider that each place has own region, such as the regions of prefectures and municipalities are different. Suppose that the region of each place as a circle, we need to measure the geographical distance between two places by considering inclusion relationships between their regions and a radius of each region. The determination of one scene of the video stream is described as follows:

$$\begin{cases} distG(i, i+1) - r(i) - r(i+1) & < \quad \alpha \\ distT(i, i+1) & < \quad \beta \end{cases}$$

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

$$\beta = \frac{\sum_{i=1}^{n-1} distT(i, i+1)}{n-1} \tag{2}$$

Function r returns a radius of a region of a place, then, $r(i)$ denotes the radius of the place i. Function $distG(i, i+1)$ returns the geographical distance between centers of two places i and $i+1$ in their appearance order, then, $distG(i, i+1)$-$r(i)$-$r(i+1)$ denotes the geographical distance between boundaries of the regions of i and $i+1$ for normalizing different regions of the places. And we set a threshold value α in Equation 2, which is an average of the geographical distance between the boundary of the region of each place and the center point c. Function $distT(i, i+1)$ returns a time width that the temporal distance between two places i and $i+1$ in their appearance order, and we set a threshold value β in Equation 2, which is an average of the temporal distance between each place and the center point. If two places satisfy the above conditions, one scene is determined by these two places.

Extracting Geographical Geometry

Since each place has own region, such as the regions of prefectures and municipalities are different, we extract geographical relationships that regional relationships between any two places in the order of their appearance in each scene of a video stream. In this work, we utilize 9-intersection [5] to extract regional relationships between two places.

Moreover, we consider overlapping boundary of the geographical regions that is not important in video composition. As shown in Table 1, we extract six patterns of geographical relationships between two places, that we unify $covers$ into $contains$, and $coveredBy$ into $inside$.

Table 1: 9-intersection and our proposed method

9-intersection	Our proposed method
$equal$	$equal$
$disjoint$	$disjoint$
$meet$	$meet$
$overlap$	$overlap$
$covers$	$contains$
$contains$	
$coveredBy$	$inside$
$inside$	

$Equal$ means that the same place name appears only one time or repetitively, i.e., Kyushu appears continuously in the video stream. $Disjoint$ means the regions of the two places are separated, i.e., Fukuoka Prefecture and Kagoshima Prefecture in Japan. In addition, we consider that there are two types of geographical distances between the regions of two places, $disjoint$-f means a long geographical distance, when it exceeds a threshold value; $disjoint$-n means a short geographical distance, when it does not exceed the threshold value. $Meet$ means the regions of two places have an overlapping boundary, i.e., Fukuoka and Saga Prefectures in Japan. $Overlap$ means the regions of the two places are overlapping, i.e., Hitoyoshi City and Kuma River in Japan. $Contains$ means the region of one place includes the region of another place of the video stream, i.e., Fukuoka Prefecture and Hakata Station in Japan. $Inside$ means the region of one place is included in the region of another place in the video stream, i.e., Hakata Station and Fukuoka Prefecture in Japan.

Extracting Semantic Structure

We extract semantic structure that intentions of video streams based on the geographical geometry of the video streams are shown in Table 2, i.e., temporal sequences and geographical relationships between given places, A and B, that appear in a video stream. When A $equal$ A, we assume that the intention of the video stream is to introduce A. When A $contains$ B, we assume that the intention of the video stream is to introduce B that belongs to A. When A $disjoint$ B and each place has its own wide region (i.e., prefectures), we assume that the intention of the video stream is to compare A and B. When A $meet$ B, we assume that the intention of the video stream is to compare A and B. When A $disjoint$ B and each place has its own narrow region (i.e., municipalities), we assume that the intention of the video stream is to describe the route between A and B. When A $overlap$ B or A $inside$ B, we assume that the intention of the video stream is to introduce the surrounding areas of A and B.

Table 2: Semantic structure based on geographical geometry

Intentions	Temporal sequences	Geographical relationships
Introduction to places	Place $A \rightarrow$ Place A	$equal$
Introduction to places	Place $A \rightarrow$ Place B	$contains$
Comparison of places	Place $A \rightarrow$ Place B	$disjoint$-n (in wide regions)
Comparison of places	Place $A \rightarrow$ Place B	$disjoint$-f (in wide regions)
Comparison of places	Place $A \rightarrow$ Place B	$meet$
Introduction to routes	Place $A \rightarrow$ Place B	$disjoint$-n (in narrow regions)
Introduction to routes	Place $A \rightarrow$ Place B	$disjoint$-f (in narrow regions)
Introduction to regions	Place $A \rightarrow$ Place B	$overlap$
Introduction to regions	Place $A \rightarrow$ Place B	$inside$

Media Synchronizing Mechanism

Prototype System

In this paper, we developed a novel interface called TV-Milan, using JavaScript. The TV-Milan synchronizes geographic contents with the TV programs on the basis of the derived semantic structure of the video streams, which aids viewers to better grasp the geographical information (i.e., tourist spots, regions, routes, etc.) when it appear in the video streams. In a big display, the interface is implemented by four parts: a video player window (left top), a list of place names (left bottom), a map window (center bottom), and a Street View (photo) window (right). A TV program plays in a video player window; simultaneously, a list of the place names and geographical geometry are extracted from the closed captions by using Scala [7]. Thus, the TV-Milan can show the TV program synchronized with a map using Google Maps API[5] in a map window, and online photos using Panoramio API[6] or Google Street View Image API[7] in the Street View (photo) window. In addition, a map or a Street View (photo) can be interactively presented on the second screen devices, e.g., tablets and smartphones.

Geographical Contents Synchronized with TV Programs
When the intention of the video stream is to introduce places, the TV-Milan presents the maps of the given places with seamless effects and the photos related to them, to help viewers easily grasp the locations and the appearances of the given places. When the intention of the video stream is to compare places, the TV-Milan presents the map, including all given places with seamless effects and their photos, to help viewers better understand the positional relationships and the appearances of the given places. When the intention of the video stream is to introduce routes, the TV-Milan presents the map and Street View for showing the routes between the given places with seamless effects, to help viewers easily know the routes between the given places. When the intention of the video stream is to introduce regions, the TV-Milan presents the map of the given places with seamless effects and the photos for showing the overlapping region of the given places, to help viewers easily grasp the locations and the atmosphere of the common region of the given places.

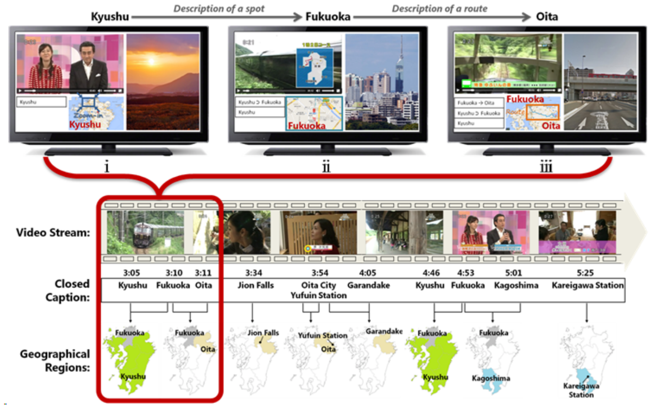

Figure 3: An example of media synchronizing mechanism

For example, suppose place names, i.e., Kyushu (3:05), Fukuoka (3:10), Oita (3:11), and so on, they appear in a video stream (see Figure 3), which is a travel program[8] introducing Kyushu in Japan, with total run-time of 23 minutes. In this case, start from the first place names appear in the video stream. When two place names are determined in one scene, TV-Milan determines the intentions of them and presents the geographic contents synchronized with the video stream. We first extract geographical geometry: Kyushu $contains$ Fukuoka, Fukuoka $disjoint$ Oita. In Figure 3, when the video stream shows Kyushu to Fukuoka in the same scene at the first time in the video player window, the map window presents the region of Kyushu and zooms-in to Fukuoka.

[5] https://developers.google.com/maps/

[6] http://www.panoramio.com/api/data/api.html

[7] https://developers.google.com/maps/documentation/streetview/

[8] http://www.nhk.or.jp/hakken/kyushu/program/tetsudo.html

Simultaneously, a Street View window first presents the photos related to Kyushu as shown in **i**. Then, when the video stream shows Fukuoka, the Street View window changes to present photos related to Fukuoka as shown in **ii**. When the video stream shows Oita after Fukuoka in the same scene, the map window presents both Fukuoka and Oita and highlights the route between them. Simultaneously, the Street View starts at Fukuoka and proceeds along the route to Oita as shown in **iii**. In this manner, TV-Milan enables viewers to easily and efficiently grasp the geographical information of places that appear in the video streams.

Conclusions

In this paper, we built a novel interface, called TV-Milan, it presents geographic contents (i.e., maps, Street View, etc.) synchronized with TV programs based on geographical geometry and semantic structure of video streams. We conducted a simple preliminary study by using actual travel programs in Japan with our developed prototype system, and the results show that TV-Milan has the potential to help users easily view the TV programs while simultaneously viewing geographic information.

In the future, we plan to improve the method for presenting geographic contents by considering cinematography and film languages. In addition, we should conduct a subjective study including more participants with different profiles (i.e., age, gender, study level, etc.). Further, we intend to expand TV-Milan to allow user interactions with the video streams for controlling the geographic contents to be presented.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26280042.

References

[1] Banjou, Y., Takakura, H., and Kambayashi, Y. Generation of rough maps according to various user requirements. In *55th National Convention of Information Processing Society of Japan (IPSJ)* (1997), 481–482.

[2] Geerts, D., Leenheer, R., D. Grooff, D., Negenman, J., and Heijstraten, S. In front of and behind the second screen: Viewer and producer perspectives on a companion app. In *Proc. of TVX2014* (2014), 95–102.

[3] Kitayama, D., Miyamoto, S., and Sumiya, K. A customizing method of digital maps based on user's operations and object's appearing patterns. *Transactions of Information Processing Society of Japan (TOD48) 3*, 4 (2010), 65–81.

[4] Kobayashi, K., Kitayama, D., and Sumiya, K. Cinematic street: Automatic street view walk-through system using modified maps. In *Proc. of W2GIS 2011* (2011), 142–158.

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

[6] Nandakumar, A., and Murray, J. Companion apps for long arc tv series: Supporting new viewers in complex storyworlds with tightly synchronized context-sensitive annotations. In *Proc. of TVX2014* (2014), 3–10.

[7] Scala. http://www.scala-lang.org/.

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

[9] Viana, W., Miron, A. D., Moisuc, B., Gensel, J., Villanova-Oliver, M., and Martin, H. Towards the semantic and context-aware management of mobile multimedia. *Multimedia Tools Appl. 53*, 2 (2011), 391–429.