Movie Map for Virtual Exploration in a City*

Kiyoharu AIZAWA (a), Fellow

SUMMARY This paper introduces our work on a Movie Map, which will enable users to explore a given city area using 360° videos. Visual exploration of a city is always needed. Nowadays, we are familiar with Google Street View (GSV) that is an interactive visual map. Despite the wide use of GSV, it provides sparse images of streets, which often confuses users and lowers user satisfaction. Forty years ago, a video-based interactive map was created—it is well-known as Aspen Movie Map. Movie Map uses videos instead of sparse images and seems to improve the user experience dramatically. However, Aspen Movie Map was based on analog technology with a huge effort and never built again. Thus, we renovate the Movie Map using state-of-the-art technology. We build a new Movie Map system with an interface for exploring cities. The system consists of four stages: acquisition, analysis, management, and interaction. After acquiring 360° videos along streets in target areas, the analysis of videos is almost automatic. Frames of the video are localized on the map, intersections are detected, and videos are segmented. Turning views at intersections are synthesized. By connecting the video segments following the specified movement in an area, we can watch a walking view along a street. The interface allows for easy exploration of a target area. It can also show virtual billboards in the view.

key words: Movie Map, street view, image processing, interaction, visual SLAM

1. Introduction

In this paper, we introduce our work on “Movie Map” by which we can virtually explore a certain area in a city.

Since last year, the corona disaster has been forcing people to avoid travel outside Japan and within Japan, which is a serious situation for many industries, especially the tourism industry. It was reported that the number of passengers on the Tokaido Shinkansen during Golden Week of 2020 was 94% less than the previous year [1]. In addition to the problems of the industry, exploration as an opportunity to learn something contributes significantly to people’s well-being [2], and this opportunity is being taken away. Therefore, there is a strong need for digital tourism and virtual exploration. In this paper, we would like to introduce an interactive video map as one of the means to realize virtual exploration.

Video archives such as YouTube have a vast amount of video footage taken by travelers in the past. These can be used as materials for digital tourism. In addition, there are attempts to go one step further and allow people to explore a particular area virtually. At present, these attempts can be roughly divided into two categories: those based on real images and those based on VR with graphics.

For a VR tool-based project, for example, the “Virtual Tokyo Tower” project can be mentioned [3]. The Virtual Tokyo Tower is built on a VR platform that allows users to visit the Tokyo Tower with their own avatars from their own rooms and enjoy communication with others, panoramic night views, events, and shopping. Thus, it is not a world of imagination like a game but a space close to reality, where interaction is built in. There are many other examples. For example, several universities have tried to use VR to visit open university campuses for the Corona disaster (e.g., Virtual University of Tokyo).

As for video, an interactive movie map can be mentioned [4, 12, 13]. A movie map is a motion picture version of Google Street View (GSV). Its origin can be traced back to the Aspen Movie Map, which was created in 1980 based on analog technology [4]. It was based on the idea of capturing images of streets and all turns at intersections in Aspen city with an on-board film camera, creating a database of these images recorded on analog optical disks, and playing them back according to interactions. The time and effort required to build the database must have been enormous, and except for the Aspen Movie Map, it has never been made. Now, a long time has passed since then, and the author’s group has made an effort to construct a movie map.
using the current advanced technology [12], [13].

In this paper, we introduce our work on our new Movie Map [13]. We consider that an image-based interactive map – GSV and Movie Map – is very effective for exploring a city. GSV uses still images, but it is highly interactive – users choose direction and views at every frame whose location is sparse. By contrast, Movie Map uses video segments, and users can set the moving speed, choose a direction at the intersection, and change viewing direction.

2. Technology for Virtual Exploration in a City - Interactive Images and Videos

2.1 Street View

Currently, Google Street View (GSV) [5] and Mapillary [7] are the popular interactive map services that present users interface that combines maps and images. Among them, GSV is widely used. It provides users with 360° images corresponding to a chosen location on the map. Users interact with these images – choose the locations on the map and the direction of the movements – and displayed images are updated in response to the movements. The view creates a realistic user experience as it enables users to experience the place as if they were there. GSV is widely used and offers convenience for tasks such as guidance and prior learning about a target area [6]. There are variants similar to GSV. For example, a work [8] presents a method building GSV type exploration for an indoor environment.

However, GSV is not perfect for virtual exploration. In GSV, street images are sparse. To view all the images along a route on GSV, users need to repeat transitions from one image to another [14]. This operation is quite tedious and occasionally difficult. Such image transitions may result in the user getting lost or being led to a wrong direction. Additionally, because of the sparsity of images, the user does not experience continuous movement.

2.2 Aspen Movie Map

The use of videos instead of sparse images can solve these problems that are experienced by GSV users. When using videos, we select the starting point and moving direction and playback a street video along the desired direction. It is possible to make video transitions at the intersections and subsequently change the direction of movement. Hence, the use of videos can eliminate the need for excessive interactions and also enable users to experience continuous movement.

Surprisingly, forty years ago, a research project prototyped Movie Map based on analog video technology, which is well known as Aspen Movie Map [4]. It was an interactive video map, built using analog optical disc technology, to engage the user in a simulated driving through the city. In this system, panoramic images were captured approximately every 10 feet (3m) by four stop-animation cameras that were placed at 90° intervals in a horizontal circle, and the locations were measured via GPS. Images along roadways and at intersections were stored in several optical disk drives. A user could interactively explore the area on a touch screen by specifying the direction of movement. The Aspen Movie Map showed a methodology for visualizing a specific area by only using video segments.

However, the technology used at that time was not sufficiently advanced, and therefore the movie map could not be scaled. Until now, minimal research in this field has been conducted – they are summarized in [10]. There are limited examples of movie maps such as [11] and [9]. [11] replay captured route movies without any route synthesis. [9] required whole manual operation of video segmentation and bifurcation. Thus, they cannot be efficiently generalized for a Movie Map to cover a given particular area.

GSV emerged in 2007. It was initially used in several cities in the US before expanding globally [5]. It provides an interactive slide-show type of view. Presently, unlike GSV, the use of the Movie Map is highly uncommon.

GSV is not considered to be necessarily better than Movie Map from the point of view of user experience. [4] claimed that Movie Map allows the user to experience an area as if they were driving through the area. Such an experience is provided to the user by using continuous visual information – this is what GSV lacks. In this regard, Movie Map is still worth studying. Therefore, we take advantage of today’s technology to renovate a Movie Map for walkers in certain city areas [12], [13]. The movie map will contribute to virtual exploration that enables us to virtually move around in a certain area in a city.

2.3 Exploration via VR

Recently, VR platforms such as Hubs [15], Cluster [16] VR-Chat [17], etc., have become readily available. By creating a CG model of the environment that closely resembles reality, we can virtually explore the environment on the VR platform. For example, Virtual Shibuya, Virtual Tokyo Tower, and Virtual Tokyo University have been created and released to the public. Users can enter the environment with their own avatar, move around freely, and interact with others in the environment.

Both VR-based and video-based systems have their advantages and disadvantages. VR-based systems can be highly interactive, but they require a vast amount of effort to build environment models. Thus the VR-based method is not scalable. Their environments are often limited to one or a few buildings. Video-based systems have limited capability in interactions. Instead, they do not require model building, and what they need is to capture street videos in the target area.

In this paper, we describe our work that builds Movie Map via videos captured along streets. The videos are processed almost automatically to create an interactive video database.
3. Movie Map Building System

In this section, we describe the processing flow of video data. Figure 2 shows the outline of our Movie Map system. The map building system is divided into three stages: acquisition, analysis, and management of data. These are further subdivided into several processes. What we need to build a Movie Map is 360° videos and two reference coordinates per video. The system automatically outputs video segments and synthesizes turning views that are used in our Movie Map database. Our system is readily applicable to various areas.

3.1 Data Acquisition

We capture 360° videos along the streets in the target area and prepare a list of coordinates of reference points.

3.1.1 Acquisition of Street Videos

360° videos are captured by persons carrying the 360° camera and walking along the streets. The camera is mounted on a pole attached to a backpack. The person carries the backpack, and the camera is positioned at a height higher than the person. We do not precisely control the height of the camera viewpoint.

We captured videos along streets in the University campus and the areas surrounding the Kyoto station, Namba station, and Yokohama Chinatown for our experiments. The overall pictures of the areas of our campus and Yokohama Chinatown are shown in Figs. 3 and 4, respectively. We physically walked in both directions to capture videos of a single street. Acquiring two-way videos is necessary for our task of producing a natural feel when moving forward and backward. It should be noted that the streets do not need to be straight, and there are no restrictions on the shape of the streets. However, we assume that two shooting streets intersect at one point. If they intersect at multiple points, we divided one of them into two paths to keep the assumption.

---

**Fig. 2** Outline of the system

**Fig. 3** One of the shooting areas (app.600x800m² of the main campus of the University of Tokyo). The lines show the streets we captured.

**Fig. 4** One of the shooting areas (app.600x500m² of Yokohama Chinatown). The lines show the streets we captured. Two lines along a street visualize both side walks.
3.1.2 Assignment of Reference Coordinates

We assign the reference point coordinates to each street video. As described in Sect. 3.2.1, we apply vSLAM to each video and estimate the relative coordinates of the camera positions of the key-frames of each video. To integrate all camera positions into a single map, we assign the map’s coordinate that is common to all captured videos.

We assign the latitude and longitude of two reference points to each video’s start and end frame, respectively. Instead of latitude and longitude, we could also use a specific coordinate defined on a map, as long as they are the same for all captured videos.

3.2 Data Analysis

In this section, we analyze the captured street videos and automatically obtain their intersection information. Establishing intersections of the streets is the most important to synthesize the route.

3.2.1 Application of vSLAM

We estimate the relative camera poses, that are, positions and orientations, by using OpenVSLAM [18] which uses visual features in all directions in the 360° image for optimization.

3.2.2 Mapping Videos onto a Common Coordinate

We map all relative camera positions onto a common coordinate space. Camera positions are independently calculated for each video by using OpenVSLAM [18]. We map all videos onto the map’s common coordinate to associate them with each other. Visual SLAM computes the camera pose in an individual world coordinate for each video. These world coordinates are not equal to each other and different from the map’s coordinate in its scale, orientation, and origin.

To align the camera positions of the video onto the map, we use the reference point coordinate information mentioned in Sect. 3.1.2. Considering the vector from the start point to the endpoint of the estimated camera positions, we obtain the rotation and scaling so that the vector equals the reference vector from the reference start point to end point coordinates.

We align the camera positions to the map by applying the translation, rotation, and scaling to the positions of the camera of the video. We also align the camera orientation to the reference vector. Repeating this process for all videos, we align all camera positions on the common coordinate of the map.

Figure 5 presents an example of this type of mapping. Red and blue markers represent the coordinates of key-frames in both ways of each street.

3.2.3 Intersection Detection

We detect the intersections by using the coordinate information of the videos and refine them by using visual features. We obtain the intersection information – which street video intersects with which video in which frame and the relative rotation between the frames of these two videos.

We find video pairs that intersect each other. Suppose video A and video B have an intersection, then we determine the most similar frames in location and visual features. We firstly search the frame pairs that are the closest in terms of the camera locations.

As shown in Fig. 5, the camera locations estimated by vSLAM contain errors. To minimize these errors, we refine the intersection frames using visual features. We optimize the intersection frame pairs by using visual similarity based on ORB features [19] for dozens of frames around the intersection detected by the location. We define the visual similarity between two frames by the sum of Hamming dis-
tances of ORB features of the top k feature point pairs that are identified matched\(^\dagger\). These refined frame pairs are the final intersection frames. By using visual similarity, the results are more accurate than those obtained when only using location information.

In the experiments, we did not capture videos in the same condition – the time changed from morning to late afternoon, and the day changes within several days. We have not tested the situation if the condition is much more different such as change over seasons.

We record the following data: the two videos to which the frames belong, the timestamp of the intersection frames in the videos, their coordinates, and the relative camera rotation between the two intersection frames.

We automatically repeat the above analysis for all pairs of streets and obtain information on all intersections in the target area.

3.3 Data Management

Using the intersection information obtained in the data analysis stage, we can synthesize a route movie by compositing the videos. To produce an interactive real-time player, we convert the videos into a format convenient to the interface. We segment the street videos into sections between intersection frames and added metadata to specify the video sections. We also synthesize the turning views at intersections, which are inserted to produce a natural transition from one video section to another.

3.3.1 Splitting Street Videos into Sections

We split the street videos at their intersections into video sections. When we use the video in our interface, the video playback unit is a section between intersections. It is significantly faster to load a section than to load the full street video. The division is performed based on the intersection frame information: the street video index and timestamp.

There are multiple sections within a single physical intersection because we maintain videos of both directions of a single street. In a typical case, as shown in Fig. 6, there are actually four video intersections in a physical intersection, and small sections are segmented, too. We add the following metadata to each section: the intersection ID and the street video ID to which it belongs.

3.3.2 Synthesis of Turning Views at Intersections

In our interface, the user can turn at the intersections. Before and after turning, we switch from one video section to another video section. However, switching movies without any interpolation makes us feel uncomfortable and causes inconsistencies in our cognition of position and direction.

In Aspen Movie Map [4], turning movies at each intersection were captured separately and inserted when the driver turned. In our case, we need eight moving patterns (four forward and four turnings) at a typical physical intersection as shown in Fig. 6. When we consider the number of intersections that exist in a certain area, shooting turning videos for all of them is not very practical. Owing to the use of 360° videos, we synthesize the turning views by rotation and blending.

Video sections before and after the intersection belong to different street videos captured at other times. Hence, it is often observed that the brightness and objects change – cars and people before and after the intersection change. Therefore, we synthesize the turning views by blending the two intersection frames that are the last and the first frames of the two video sections. We set the time length of the turning view to one second.

We synthesize all the possible turning views in advance and store them in the movie map database. Then, when using the movie map, video segments and turn views along the user-directed route are played back on the fly, switching between them.

4. Movie Map Exploring Interface

We propose an exploration-type interface based on the Movie Map made of 360° street videos analyzed and managed as described in the previous sections. This interface is helpful for purposes of prior learning and a virtual tour of a target area. Figure 1 shows an overall view of the interface. On this interface, the user can see street videos that correspond to the coordinates on the map. By selecting the desired direction at an intersection, we can easily manipulate the directions of movement. An example of the route switching movie at an intersection is shown as a subsampled slide show in Fig. 7.

4.1 Interaction on the Interface

Based on the metadata of the intersections, the intersection points are marked on the map. We can start exploration by choosing one of the intersection points.
AIZAWA: MOVIE MAP FOR VIRTUAL EXPLORATION IN A CITY

Fig. 7 An example of a sequence of a synthesized route movie. The top left is the last frame before switching, and the bottom right is the first frame after the switching.

Fig. 8 Select start point to explore

Figure 8 is another starting screen for determining the starting point of this exploration. Landmarks in the target area are defined as the starting locations in advance, and arrows indicating possible walking directions are displayed. The user selects one of the landmarks and a walking direction to start the exploration. When the user starts exploring, a screen of a walking video appears, and the video starts playing.

The user can change the viewing direction by dragging the 360° video and can easily specify the walking direction at the intersections by choosing the arrow that appears when approaching an intersection as shown in Fig. 8. By clicking on one of the arrows, the user can select the desired direction, turn the view accordingly, and switch to the next video section. The user can change the location any time by clicking an intersection point on a map. In addition, the user can perform basic video manipulations such as change of playback speed by interacting with the button inputs.

In this way, the user can walk continuously in a chosen area at their desired speed. Furthermore, without the user selecting directions, the system can autonomously produce moving video by random selection of the direction.

4.2 Hiding Details of Intersections for Visualization

In the interface shown in Fig. 1, the intersection on the map is visualized as a single point. The actual data, as described in Sect. 3.1.1, consists of two-way videos of a street – one physical intersection generally contains four intersection frame pairs. We group the multiple video intersections of a single physical intersection as a point and draw its cluster center as the point on the map.

We keep the connections of the grouped intersections as a graph structure. Based on this data, when approaching an intersection, the system displays the navigation arrows only for the possible directions of walking. The system visualizes the current position on the map using the frame information of the playback video – the position moves along the street as the video plays.

4.3 Virtual Billboards

We can display virtual billboards in the view. After building the Movie Map, the video is aligned in the direction of the street. We can overlay a billboard in the video by corresponding it to its specific location, as shown in Fig. 9. The billboard can appear at varying distant viewpoints. A location per each billboard is maintained in a billboard list. We use the timestamp of the video to specify their location. When the user approaches the location, the corresponding billboard pops up. Each billboard can be associated with additional information such as HP, which appears by clicking the virtual billboards.
4.4 Prototype

We prototyped our Movie Map at [20]. We can interactively select the directions and play route movies composited on the fly. The demonstrations of the University campus and Yokohama china town are available†. The screenshots of the demonstrations of the University Campus and Yokohama china town are shown in Fig. 1 and Fig. 10, respectively. A demonstration of a certain area around Kyoto Station captured is shown at [21].

5. Conclusion

We present our new Movie Map system and its exploration interface. The proposed method involved the acquisition, analysis, management, and interaction of 360° videos. Once we acquired the street videos, the entire pipeline of processing was almost automated.

Our system segmented videos into sections using the information of the detected intersections, and the sections were refined by using visual similarity. Moreover, intersection turning views were synthesized in advance. The walking videos of the target area were displayed by compositing video sections and turning views according to the route interactively specified by the user. The manipulation of the interface is of a simple selection of the directions at intersections. With additional operations, we could display virtual billboards. We prototyped a system that allows us to virtually explore the University campus and Yokohama china town.

Acknowledgements

We thank Mr. Sugimoto and other students of our laboratory and members of VTEC Laboratories Inc., who contributed to this Movie Map project. A part of this work was supported by VTEC Laboratories Inc.

References

[1] Nihon Keizai Shimbun, Use of Tokaido Shinkansen down 90% during Golden Week as people refrain from going out, May 7 2020. (in Japanese)

[2] B.Q. Ho and T. Abe, “Escape From Usual Life: The Tourist Gaze and Well-Being,” Japanese Journal of Serviceology, vol.6, no.1, pp.20–27, 2019. (in Japanese)

[3] Virtual Tokyo Tower Official HP https://virtual-tokytower.com/

[4] A. Lippman, “Movie-maps: An Application of the Optical Videodisc to Computer Graphics,” Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’80). ACM, pp.32–42, 1980.

[5] Google. https://www.google.com/intl/en/streetview/explore/ 2005.

[6] A.G. Rundle, M.D.M. Bader, C.A. Richards, K.M. Necker-man, and J.O. Teitler, “Using Google Street View to audit neighborhood environments,” American journal of preventive medicine, vol.40, no.1, pp.94–100, 2011.

[7] Mapillary AB. Mapillary. https://www.mapillary.com/ 2014.

[8] O. Saurer, F. Fraundorfer, and Marc Pollefeys “OmniTour: Semi-Automatic Generation of Interactive Virtual Tours from Omnidirectional Video,” Proc. 3DPVT2010 (Int. Symp. 3D Data Processing, Visualization and Transmission), 8pages, 2010.

[9] M. Uyttendaele, A. Criminisi, S.B. Kang, S. Winder, R. Szeliski, and R. Hartley, “Image-Based Interactive Exploration of Real-World Environments,” IEEE Comput. Graph. Appl. vol.24, no.3, pp.52–63, 2004.

[10] M. Naimark, A 3D Moviemap and a 3D Panorama, vol.3012, Proc. SPIE, 2004.

[11] M. Naimark, Field Recording Studies, Immersed in Technology: Art and Virtual Environments, 299-302, M.A. Moser, ed., MIT Press, 1996.

[12] N. Sugimoto, T. Okubo, and K. Aizawa, “Urban Movie Map for Walkers: Route View Synthesis using 360 degree Videos,” Proceedings of 2020 International Conference on Multimedia Retrieval, pp.502–508, 2020.

[13] N. Sugimoto, Y. Ebine, and K. Aizawa, “Building Movie Map - A Tool for Exploring Areas in a City - and its Evaluations,” Proceedings of the 28th ACM International Conference on Multimedia, pp.3330–3338, 2020.

[14] K. Kobayashi, D. Kiyayama, and K. Sumiya, An Automatic Street-View Generating System based on User’s Digital Map Operation, 2010.

[15] https://hubs.mozilla.com/

[16] https://cluster.mu/

[17] https://hello.vrchat.com/

[18] S. Sumikura, M. Shibuya, and K. Sakurada, “OpenVSLAM: A Versatile Visual SLAM Framework,” Proceedings of the 27th ACM Multimdia, pp.2292–2295, 2019.

[19] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “ORB: An efficient alternative to SIFT or SURF,” International Conference on Computer Vision, pp.2564–2571, 2011.

[20] Our Movie Map Prototype Demo site of The University Campus and Yokohama China Town. https://moviemap.jp (Hongo Campus is currently limited to the University internal network, but Yokohama China Town is widely available.)

[21] Movie Map demonstration Youtube https://www.youtube.com/watch?v=l_Dnul6QjiQ
Kiyoharu Aizawa received the B.E., the M.E., and the Dr.Eng. degrees in Electrical Engineering all from the University of Tokyo, in 1983, 1985, 1988, respectively. He is currently a Professor at Department of Information and Communication Engineering of the University of Tokyo. He was a Visiting Assistant Professor at University of Illinois from 1990 to 1992. His research interest is in multimedia applications, image processing and computer vision. He received the 1987 Young Engineer Award and the 1990, 1998 Best Paper Awards, the 1991 Achievement Award, 1999 Electronics Society Award from IEICE Japan, and the 1998 Fujio Frontier Award, the 2002 and 2009 Best Paper Award, and 2013 Achievement award from ITE Japan. He received the IBM Japan Science Prize in 2002. He is on Editorial Boards of IEEE MultiMedia, ACM TOMM, APSIPA Transactions on Signal and Information Processing, and International Journal of Multimedia Information Retrieval. He served as the Editor in Chief of Journal of ITE Japan, an Associate Editor of IEEE Trans. Image Processing, IEEE Trans. CSVT and IEEE Trans. Multimedia. He was a president of ITE and ISS society of IEICE, 2019 and 2018, respectively. He has served a number of international and domestic conferences; he was a General co-Chair of ACM Multimedia 2012 and ACM ICMR2018. He is a Fellow of IEEE, IEICE, ITE and a council member of Science Council of Japan.