GENERATION OF VIDEO METADATA SUPPORTING VIDEO-GIS INTEGRATION

In-Hak Joo, Tae-Hyun Hwang, and Kyung-Ho Choi

Electronics and Telecommunications Research Institute
Spatial Information Technology Center
161 Gajeong-dong, Yuseong-Gu, Daejeon, 305-350, KOREA

ABSTRACT

Geospatial information expressed by video can provide more realistic and comprehensible information that cannot be obtained from digital map. We introduce video and integrate it to GIS by supporting video-map cross reference and bi-directional search. To support content-based search of geospatial objects appearing on video, we generate video metadata that has necessary information of geospatial objects. The most important part of the video metadata is outline of object on video frame. The outline of an object varies frame to frame, and thus should be generated for every frame in which the object appears. Because the object moves rapidly between video frames collected with 1-second interval, object tracking cannot be easily done by simple algorithm. In this paper, we devise a semi-automatic object tracking method by combining photogrammetric solution with image-based method.

1. INTRODUCTION

As GIS (Geographic Information System) is spread widely, the needs for more realistic information of geospatial data are increasing. However, most of conventional GISs mainly provide simple information expressed by map. Because such GIS is not so informative to human, many researches are being conducted about introduction of multimedia in the field of GIS. Especially, information expressed by video, in nature, can provide realistic and comprehensible information of geospatial object that cannot be obtained from simple map. For this reason, video has been introduced and used as useful information in the field of GIS [1]. The GIS supported by video information is spotlighted as a new approach that can overcome the weakness of conventional map-based GIS. However, up to present, video is handled just as an attribute of geospatial object, and relationship between map and video is established restrictively, manually, and inaccurately. Such GIS cannot support content-based search of geospatial object on video.

To solve the problem, video-GIS integration technology is suggested where geospatial information is represented and searchable on video. The geospatial objects appearing in video frame are constructed and represented as video metadata [2], which supports visual browsing and content-based search of geospatial information on video. The most important part of the video metadata is outline of object, but its generation method is quite difficult job. Therefore, in this paper, we suggest a semi-automatic object tracking method for the generation of outline of objects in video.

2. VIDEO-GIS INTEGRATION

In this section, we discuss about issues of integration of video and GIS. The main functions of video-based GIS are (1) to construct geospatial database from video; and (2) to generate video-map cross reference and support bi-directional search. We briefly discuss the first topic, and concentrate on the second topic.

There are some studies about construction of geospatial data from video or image sequence, including [1]. The geospatial database can be constructed by using vehicle-borne video and corresponding information of camera, without expensive actual field survey. Collection of video data and camera information should be done synchronized and accurately, so sophisticated sensors and hardwares are integrated to mobile mapping system. We use a mobile mapping system called 4S-Van [3] as a video-collecting platform for our subsequent discussions. Detailed explanation of mobile mapping system is summarized in [3].

2.1. Video-Map Referencing

Although video can provide realistic and human-perceptible information of geospatial objects, we cannot get information of geospatial objects from it, because video has no explicit geospatial information by itself. Therefore it is necessary to generate explicit reference between video and map that can provide bi-directional search and browsing.

When users who are interested in geospatial objects browse video or digital map, the most frequent types of queries are ”Find the video frame where the city hall appears” (map-to-video query), and ”What is the location of object that I am looking now on current frame?” (video-to-map query). So

0-7803-8554-3/04/$20.00 ©2004 IEEE.

the data for geospatial object that should be provided are its location in pixel coordinate, location in actual world coordinate, and important attributes. Among the data, location in pixel coordinate is most important because it is a key to maintain reference between two media. It is represented by outline surrounding the object, usually set of feature points. Other data such as actual world coordinate and attributes are existing ones, usually stored in geospatial database. How to get the data is simple and a matter of implementation. We use the outline as a key to support indexing geospatial objects on video.

2.2. Video Metadata

To express information of objects appearing in video, the description about the information is encoded as video metadata. We support video-map cross referencing by generating the video metadata, and design an indexing scheme [2]. For general use of video data in various applications, the metadata should be structured and written by standardized format. The most well-known and widely used format for description of video metadata is XML-based MPEG-7. The design of video metadata for description of geospatial information is suggested in [2]. The video metadata includes frame information that contains information of objects therein, such as geo-coordinate, outline, keywords, and attributes.

3. GENERATING VIDEO METADATA

As mentioned in section 2.1, the outline is the most important data that we should generate for the video metadata that supports content-based search for geospatial objects. The outline of an object varies frame to frame, and thus should be generated for every frame in which the object appears. Because an object can appear in many frames and there are many objects in video, the generation of outline is time-consuming job if it is done manually. We suggest a method for generating the outlines of objects in semi-automatic manner.

3.1. Generating Outline by Object Tracking

To generate the outline of objects, image processing technologies such as object tracking or segmentation are necessary. There are many object tracking methods suggested based on image processing [4], and most widely used object tracking method is block matching, which analyzes characteristics of pixel value to track objects [5].

In our work, the video is collected by vehicle running along the road. It has characteristics that (1) non-interested objects such as pedestrians, vehicles, and backgrounds are moving as well as interested objects such as buildings and roadside facilities; (2) the objects may not be distinguished from background clearly; and (3) video is collected with 1 frame/sec frame rate, so the object moves rapidly from a frame to next frame.

We suggest an object tracking method for the data with such characteristics. Especially, the third condition is critical factor that makes object tracking difficult. If we consider the video with high frame rate (about 20 frame/sec), the offset of object region between adjacent frames is small and object tracking becomes simple. However, because we handle video with low frame rate where object tracking is difficult, we suggest a complementary method for such video in section 3.2.

3.2. Phtotogrammetry-aided Object Tracking

When video is collected with 1 frame/sec frame rate, object moves rapidly from a frame to next frame. In case of our experimental video with 1392*1040 size, pixel distance of the same object between two adjacent frames is about 150∼160. Such video cannot be easily handled by simple object tracking solution, because it requires large search area that means longer processing time and higher possibility of failure. To solve the problem, we devise a semi-automatic method with photogrammetric solution combined to image-based object tracking.

With photogrammetric method used, we can estimate location of an object in video frame as long as the world coordinate of the object is known [6]. Note that, stereo image is required for the photogrammetry calculation. When the video or image sequence is collected, the exact position and attitude of camera synchronized with each frame are also collected. With the camera information, we can calculate 3-dimensional world coordinate of an object from the pixel coordinate of the object [6]. We call the function *Intersection*. Again, the 3-dimensional world coordinate of the object can be converted to pixel coordinate of any frame, as long as the object is visible in the frame and the camera information is provided for stereo images [6]. We call the reverse conversion *Resection*. These two functions are essential functions of photogrammetry.

The two functions are applied to a feature point of an object as follows. In the frame in which an object first appears, we convert its feature point (it may require user input) to world coordinate. For the next frame, the world coordinate is converted to pixel coordinate, which is the estimated pixel corresponding to the feature point.

Because the estimated pixel may not be exact, we do block matching operation near the estimated pixel to find actual feature point. Mask window size for block matching is 32*32, and search window size is 32*64 because the displacement of feature point may be caused by vehicles's vertical vibration. We select ZSAD [5] as the metric for block matching. When determining the mask window size, we consider that (1) the feature point should be input at two

stereo image and may be inaccurately input by user; (2) *Intersection* and *Resection* may yield inacurracy, depending on several factors (our previous experiments [3] show that the error is below $50cm$ in world coordinate and below 30 pixels in pixel coordinate); and (3) too large search window may bring inaccurate result. Other parameters and thresholds for the block matching process are empirically determined, mainly depending on data characteristics and accuracy of photogrammetry operations. Once actual feature point is found, such process continues for the following frames until the object disappears.

3.3. Adjusting Mask Window Size

In vehicle-borne video, the size of object in video frame is increasing as the vehicle goes forward. If user inputs feature point of an object at first frame in which the object appears, the feature point appears small, which is likely to cause inaccuracy. Therefore, we reverse the processing order, i.e. user inputs the feature point in the latter frame and applies object tracking backward with regard to time.

Further, because the size of object in video frame is decreased in the next frame, we also decrease mask window size by resampling original block around feature point when the feature point is estimated and found in the next frame. The decreasing rate of mask window size is determined in proportion to increasing rate of distance between camera and object (in world coordinate), which can be calculated from camera parameters, position, and position of object. By using such adjusted mask window size, the processing time of block matching becomes shorter, and the method becomes more likely to success to find actual point.

3.4. Detection of Object by Segmentation

Once the feature point is estimated and found, the object can be found by object segmentation with the point used as seed point. We use region growing method that expands region, from a given seed point, by adding similar neighbor pixels [7]. The similarity may be determined by brightness, color, texture, or their combination. In our experiments, gray-level brightness is used as the criteria for region growing method. Before the object segmentation, shrink, erosion, and dilation operation are executed sequentially in order for the segmentation operation to yield better result. Note that, although many object segmentation methods can be applied to our work, their comparisons are beyond the scope of this paper.

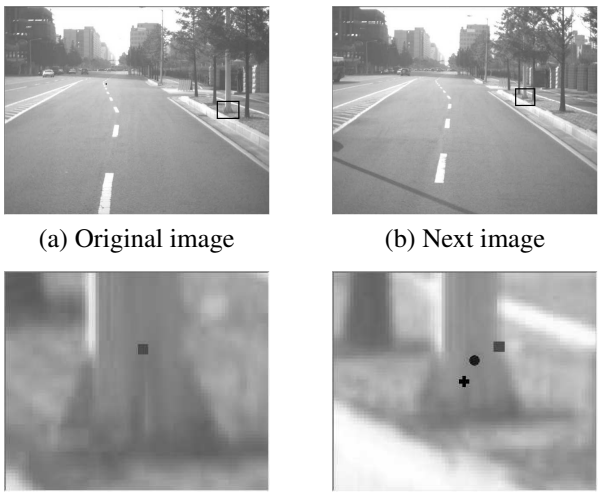
4. EXPERIMENTAL RESULTS

We experiment the suggested method with sample video with 1392*1040 resolution, 1 frame/sec frame rate, and 2-minute length, collected by running vehicle at Daejeon.

4.1. Block Matching

The result of block matching is shown in Fig. 1. The mask window size is 32*32. In the next frame, we experiment two cases: fixed-sized mask window and variable-sized mask window. Fig. 1(a) is the original image, and Fig. 1(b) is next image (with regard to time, 1 second before). Fig. 1(c) is feature point that user input at image Fig. 1(a)(input should be done in both left and right images; one is omitted). Fig. 1(d) represents the estimated feature point on next image Fig. 1(b). The point marked as rectangle is estimated feature point using photogrammetry; the point marked as cross is result of block matching with fixed-sized mask window (32*32); and the point marked as circle is result of block matching with variable-sized mask window (22*22).

We can see that the object becomes smaller, and that the block matching with variable-sized mask finds more accurate actual point. It also outperforms block matching with fixed-sized mask in processing time, proportional to mask window size. In our experiment, the mask window size is reduced by about 70% in the adjacent frame, and so is the processing time. Table 1 shows the accuracy of block matching for each block size. The compared value is Root Mean Square(RMS) error between manually determined coordinate and automatically extracted coordinate. The lower RMS value means the better accuracy.



(a) Original image (b) Next image

(c) Input point (d) Estimated point

Fig. 1. Block matching results.

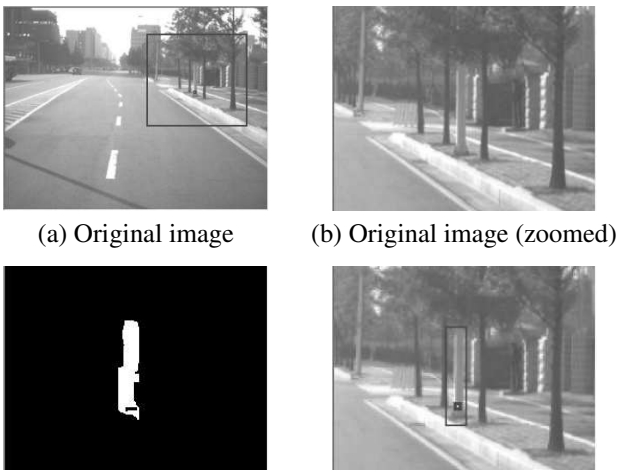
4.2. Object Segmentation

We apply object segmentation for the feature point found by block matching, to find the shape of the object. The original image and feature point is shown in Fig. 2(a) and 2(b). Fig. 2(c) shows result of segmentation, and Fig. 2(d) shows

Block Sampling Scale (Block Size)		Frame 1 (User Input)	Frame 2	Frame 3	Frame 4
Fixed Size	1.0 (32x32)	0	1.0656	1.8592	3.1223
	0.8 (26x26)	0	1.08	1.7462	3.0314
	0.6 (19x19)	0	0.05	0.0674	0.2386
	0.4 (13x13)	0	0.907	3.8799	53.6992
Variable Size		0	0	0.0586	0.2021

Table 1. Comparison of accuracy of block matching.

found region simplified as rectangle. This region is final result of our suggested method and will be recorded as outline of the object in video metadata.



(a) Original image (b) Original image (zoomed)

(c) After segmentation (d) Result (on original image)

Fig. 2. Segmentation results.

5. CONCLUSION

In this paper, we suggest a video metadata to support video-map integration, and devise semi-automatic object tracking method with photogrammetric solution combined for generating outline of objects appearing on video frame.

Because information expressed by video provides realistic information of geospatial objects, we integrate video with digital map and establish cross reference to support bi-directional search. To support content-based search of geospatial objects appearing on video, the video metadata is necessary. The most important part of the video metadata is outline of objects on video frame, so we suggest semi-automatic generation method of outline of objects. The method is based on object tracking by block matching and object segmentation. However, because such image processing methods are likely to fail for video with 1 frame/sec frame rate, we suggest a new method by applying photogrammetric solution to complement object tracking. Also suggested is adjusting mask window size of block matching for the enhancement of performance. We experiment the suggested method with vehicle-borne video to verify our method.

Though the suggested method yields good result, its performance is sensitive to accuracy of photogrammetric operation and quality of collected data. Our approach fails if the estimated feature point is too far from actual pixel (about 64 or higher in pixel coordinate). More sophisticated method to solve the problem is required in near future.

6. REFERENCES

[1] Jae-Jun Yoo, In-Hak Joo, Kwang-Woo Nam, and Jong-Hoon Lee, "The design and implementation of video geographic information system," *Proc. 29th KISS Fall Conference*, pp. 274–276, Oct 2002.

[2] In-Hak Joo, Tae-Hyun Hwang, Kyoung-Ho Choi, and Byung-Tae Jang, "A generation method of spatially encoded video data for geographic information systems," *Proc. ISRS(International Symposium on Remote Sensing)*, vol. 2, pp. 29–31, Nov. 2003.

[3] Seung-Yong Lee, Seong-Baek Kim, Ji-Hoon Choi, and Jong-Hun Lee, "4S-Van: A prototype mobile mapping system for GIS," *Korean Journal of Remote Sensing*, vol. 19(1), pp. 91–97, 2003.

[4] F. Marques and C. Molina, "Object tracking for content-based functionalities," *SPIE Vis. Commun. Image Processing*, vol. 3024, pp. 190–199, Feb. 1997.

[5] A. Giachetti, "Matching techniques to compute image motion," *Image and Vision Computing*, vol. 18, pp. 247–260, 2000.

[6] Paul R. Wolf and Bon A. Dewitt, *Elements of Photogrammetry: With applications in GIS*, 2000.

[7] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 16(6), pp. 641–647, Jun. 1994.