Object Tracking for a Video Sequence from a Moving Vehicle: A Multi-modal Approach

Tae-Hyun Hwang, Seong-Ick Cho, Jong-Hyun Park, and Kyoung-Ho Choi

ABSTRACT—This letter presents a multi-modal approach to tracking geographic objects such as buildings and road signs in a video sequence recorded from a moving vehicle. In the proposed approach, photogrammetric techniques are successfully combined with conventional tracking methods. More specifically, photogrammetry combined with positioning technologies is used to obtain 3-D coordinates of chosen geographic objects, providing a search area for conventional feature trackers. In addition, we present an adaptive window decision scheme based on the distance between chosen objects and a moving vehicle. Experimental results are provided to show the robustness of the proposed approach.

Keywords—Object tracking, mobile mapping system, photogrammetry.

I. Introduction

Emerging location-aware services including a cyber tour guide [1] and car navigation system [2] are becoming popular because they can provide more realistic information of the surrounding environment. By combining geo-referenced video with traditional geographic tools such as a map, and supporting a bi-directional search between video and geographic objects, for example, city hall, library, and so on, more effective location-aware services can be provided to mobile users [3], [4]. To provide a bi-directional search between video and geographic objects within a large area, a moving vehicle can be used to capture video sequences along roads, while a robust object tracker is essential to make indices between objects in the video sequences and the corresponding objects on a map. However, due to the movement of a vehicle, external camera parameters, for example, the location of the camera center and viewing angle, are changed, lowering scene dependency between consecutive frames. This makes traditional region-based [5], feature-based [6], and model-based [7] tracking approaches fail to find the corresponding feature points in consecutive frames. In this letter, we present a multi-modal approach for tracking geographic objects in a video sequence recorded from a moving vehicle equipped with a global positioning system (GPS), inertial navigation system (INS), and charge-coupled device (CCD) cameras.

The organization of this letter is as follows. In section II, the proposed multi-modal object tracking scheme is presented. In section III, experimental results including an adaptive window scheme are described. Finally, conclusions are given in section IV.

II. Multi-modal Object Tracking Scheme

In object tracking, although there is the basic assumption that the tracking vehicle moves forward without rapid speed changes, it is still a tough problem to track geographic objects for video sequences recorded from a moving vehicle. Due to the movement of the vehicle, conventional feature-based and region-based tracking algorithms can easily fail to track geographic objects because of low scene-dependency between consecutive frames. In this letter, we propose a novel multi-modal tracking scheme to overcome the problem. More specifically, we integrate conventional vision-based techniques into a multi-modal framework to take advantage of positioning and attitude information of the moving vehicle obtained from a GPS or INS, providing an initial search area to track geographic objects more robustly. In our approach, 3-D coordinates of chosen feature points are chosen from an

Manuscript received June 28, 2005; revised Mar. 06, 2006.

Tae-Hyun Hwang (phone: +82 42 860 1577, email: hth63339@etri.re.kr), Seong-Ick Cho (email: chosi@etri.re.kr), and Jong-Hyun Park (email: jhp@etri.re.kr) are with Telematics & USN Research Division, ETRI, Daejeon, Korea.

Kyoung-Ho Choi (email: khchoi@mokpo.ac.kr) is with the Department of Electronics Engineering, Mokpo National University, Mokpo, Korea.

intersection, and corresponding image coordinates in the following frame are estimated through a *resection*. *Resection* is a process to calculate a pixel coordinate on a selected image by using a camera location from a GPS, camera angles in three dimensions, that is., camera attitude, and camera parameters such as focal length, principle coordinate, and distortion. *Intersection* is the reverse process of *resection*. The ground location of the chosen point is calculated by using two points representing the same point in the stereo images, location and attitude information of two cameras, and camera parameters. *Intersection* and *resection* are defined as

$$P_w(X, Y, Z) = Intersection(p_l, p_r, C_l, C_r, A_l, A_r) \quad (1)$$

$$p\,(x, y) = Resection(P_w, C_l, C_r, A_l, A_r), \quad (2)$$

where P_w denotes the world coordinate, p_l and p_r indicate pixel coordinates from the left and right images, C is the 3-D position of the cameras, and A denotes their attitude [8]. Figure 1 shows a block diagram for the proposed scheme. The whole procedure can be described as follows:

Step 1. 3-D coordinate and attitude of the vehicle are obtained from a GPS or INS at time t.

Step 2. A feature point $p(x_t, y_t)$ is chosen from an image sequence by user-input at time t, and its 3-D coordinate is calculated using *intersection*.

Step 3. In the following frame at time t-1, with the known 3-D coordinate and attitude of the vehicle from a GPS or INS, the image coordinate $p(x_{t-1}, y_{t-1})$ of the corresponding feature point can be estimated by *resection*.

Step 4. The estimated image coordinate $p(x_{t-1}, y_{t-1})$ is fed into a feature tracker as an initial guess. Within the feature tracker, a region-based [5] or feature-based [6] tracking algorithm can be used to find the corresponding feature point.

Step 5. *Resection* and the feature tracker are called again to track the chosen feature point in the following frames, combining data from a GPS or INS in the proposed multi-modal tracking scheme.

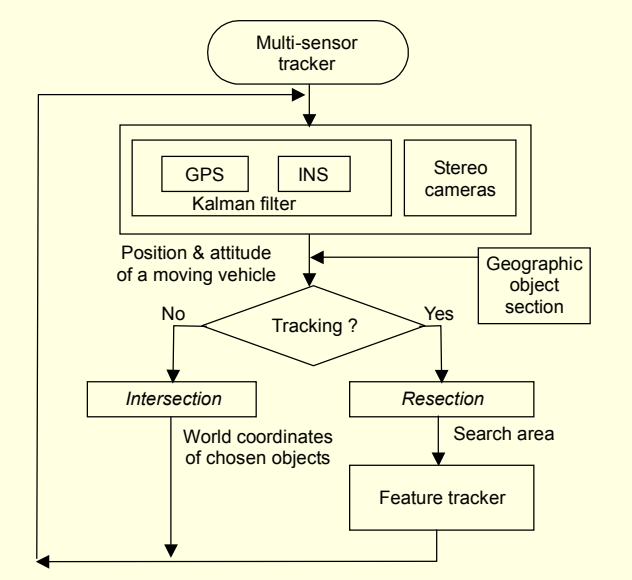

Fig. 1. A block diagram of the proposed multi-modal tracking scheme.

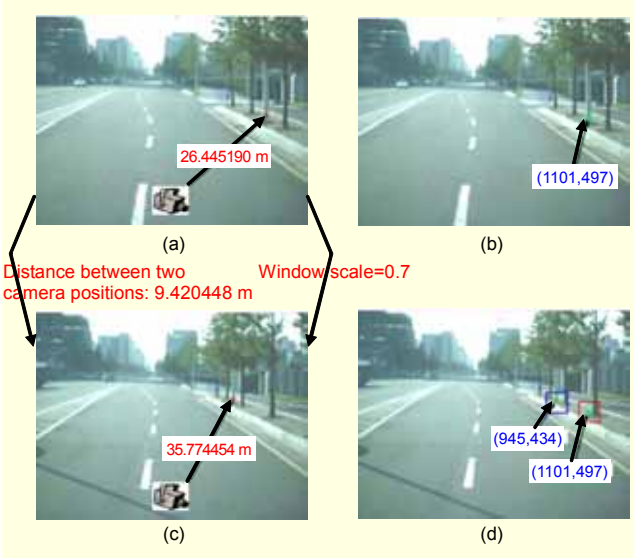

Fig. 2. An example of choosing a window size and comparing search area. (a), (b): Frame n; (c), (d): Frame n-1.

The proposed tracking approach is performed in reverse time order due to the fact that 1) images are captured as the vehicle is moving forward, 2) an object in an image captured at time t is smaller than the same one captured at time t+1, and 3) an image in which a large-enough object was captured should be chosen for better tracking performance.

III. Experimental Results

For the experiments, video sequences were collected from a moving vehicle equipped with a GPS, INS, and CCD cameras in downtown Daejeon, South Korea. The speed of the moving vehicle was around 30 km/h to capture as many geographical objects such as street lamps and traffic lights as possible. In a typical scenario, where a bi-directional search between a map and video sequences is provided, vehicle speed should not be too fast in order to provide high-quality video sequences for chosen geographic objects in a map. The video sequences were captured at the rate of 1 frame per second, with a 1300×1024 resolution using a total of 1000 recorded frames.

In the experiments, we would like to show the difficulty of object tracking in a video sequence recorded from a moving vehicle and provide the preliminary experimental results of the proposed idea. There are two issues that we have to take care of.

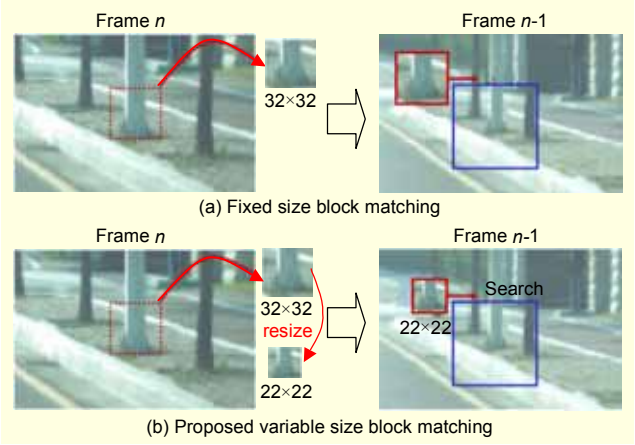

Fig. 3. The mask sizes of fixed size block matching and the proposed variable-size block matching.

First, the position of an object is changed significantly due to the movement of the vehicle. Figure 2 shows two images recorded at time *n* and time *n*-1. Note that a geographical object, marked as a green point in Fig. 2(b), is located at (1101, 497) in a pixel coordinate at the *n*-th frame. To find the corresponding point in the next frame, a search area can be defined as shown in Fig. 2(d). The red rectangular area drawn around the coordinate of (1101, 497), where the green point was located, can be a search area. However, due to the movement of the vehicle, the corresponding point is out of the search area—the point actually was moved to (945, 434) at the next frame—which means we cannot find the corresponding point in the provided search area. Second, the size of an object can be changed depending on the speed of the vehicle and video capturing rate. As shown in Figs. 2(b) and 2(d), the size of a street lamp changed too significantly to use a block-matching algorithm. To solve the problem, we present the multi-modal tracking scheme. First, we provide a search area by combining *intersection* and *resection* techniques. The blue rectangular area shown in Fig. 2(d) is the search area provided in the proposed scheme, which contains the corresponding point chosen at frame n. Second, we present a variable-size block-matching algorithm. Figure 3 shows the magnified version of Fig. 2(d) around the search area chosen in the proposed approach. The figure shows block-matching algorithms using a 32 × 32 window, shown in Fig. 3(a), and a 22 × 22 window, shown in Fig. 3(b). To minimize the effects that come from the different sizes of the chosen objects between two consecutive frames in the block-matching algorithm, we present a variable-size block-matching algorithm.

Table 1 shows a performance analysis of geographic object tracking for a blocking matching algorithm with a fixed window size, for example, 16 × 16, 22 × 22, and 32 × 32, combined with *intersection* and *resection*. The root-mean-square

Table 1. Performance analysis of multi-modal object tracking over consecutive frames.

	RMS error (pixel)			
Block sampling scale	Frame *n*	Frame *n*-1	Frame *n*-2	Frame *n*-3
1.0 (32×32)	0	3.840264155	4.608531656	7.910810842
0.7 (22×22)	0	3.884852505	4.092122199	7.198721157
0.5 (16×16)	0	3.788073521	4.601048326	5.291161134
Variable scale	0	2.254346754	3.153494186	3.370449578

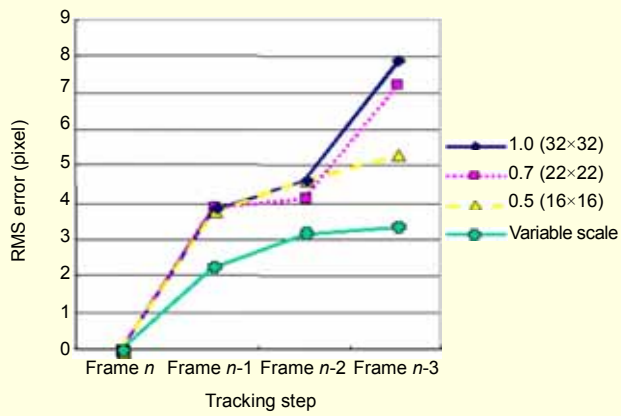

Fig. 4. Error rate of object tracking for various window types.

(RMS) error indicates Euclidian errors between a ground-truth point (x_{truth}, y_{truth}) and the tracked point in the next frame (x_t, y_t) in pixel coordinates. Although *intersection* and *resection* provided an initial search area, the performance of object tracking decreased gradually over time due to the changes of size for the chosen objects. This performance degradation was avoided by changing the window size adaptively. The window size was decided upon by calculating the distance between chosen objects and the moving vehicle. As the distance gets bigger, the window size becomes smaller. Figures 2(a) and 2(c) show the procedure to choose the window size. For instance, the distance between the chosen object and the camera is 26.4 meters at frame *n*, and becomes 35.7 meters at frame *n*-1, which means the object becomes smaller. The window is downsized to a scale of 0.7 (=26.4/35.7) depending on the distances. In our implementation, a 32 × 32 window was used for a reference block. Then, depending on the calculated size of the window, the reference block was re-sampled and used to find the corresponding block in the next frame. Figure 4 shows a comparison of the average performances of object tracking for several fixed windows and a variable window. As can be seen in Fig. 4, the variable window approach outperformed other fixed window methods, producing low errors in consecutive frames.

IV. Conclusions

In this letter, we proposed a multi-modal tracking scheme that can be applied to track geographic objects from a video sequence collected from a moving vehicle. By taking advantage of position and attitude information of a moving vehicle, calculated from a GPS or INS, a search area can be provided for object tracking in the proposed scheme. We also presented an adaptive window scheme for a conventional block-matching algorithm. Our preliminary experimental results show that the proposed scheme can track geographic objects robustly.

References

[1] G. D. Abowd, C.G. Atkeson, J. Hong, S. Long, R. Kooper, and M. Pinkerton, "Cyberguide: A Mobile Context-Aware Tour Guide," *Wireless Networks,* vol. 3, 1997, pp. 421-433.

[2] Tomoyasu Nakatsuru, Yasuyoshi Yokokohji, Daisuke Eto, and Tsuneo Yoshikawa, "Image Overlay on Optical See-Through Displays for Vehicle Navigation," *Proc. IEEE and ACM Int. Sym. on Mixed and Augmented Reality*, 2003, pp. 286-287.

[3] Toni Navarrete and Josep Blat, "VideoGIS: Segmenting and Indexing Video Based on Geographic Information," *AGILE Conf. Geographic Information Science*, 2002, pp. 1-9.

[4] In-Hak Joo, Tae-Hyun Hwang, and Kyung-Ho Cho "Generation of Video Metadata Supporting Video-GIS Integration," *Proc. Int'l Conf. Image Processing (ICIP) 2004*, vol. 3, Oct. 2004, pp.1695-1698.

[5] Changick Kim and Jenq-Neng Hwang "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 2, Feb. 2002, pp. 122-129.

[6] H. T. Tsui, Z. Y. Zhang, and S. H. Kong "Feature Tracking from an Image Sequence Using Geometric Invariants," *Proc. IEEE Int. Con. on Computer Vision and Pattern Recognition*, 1997, pp. 244-249.

[7] Y. Yoon, A. Kosaka, J. B. Park, and A. C. Kak, "A New Approach to the Use of Edge Extremities for Model-Based Object Tracking," *Proc. IEEE Int'l Conf. on Robotics and Automation*, April 2005, pp. 1883-1889.

[8] Edward M. Mikhail, James S. Bethel, and J. Chris McGlone, *Introduction to Modern Photogrammetry*, Wiley, 2001.