Reference-Frame Selection at Walking Navigation Based on Prerecorded Video

Kazuho Kamasaka †
University of Tsukuba

Itaru Kitahara ‡
University of Tsukuba

Yoshinari Kameda ‡
University of Tsukuba

Yuichi Ohta ‡
University of Tsukuba

ABSTRACT

On urban street walking navigation, accuracy of image retrieval plays an important role in our navigation system [1]. The accuracy is affected by the images in the image database. We propose a new method of selecting images from a pre-recorded video that was taken along the path people may want to trace. Our system has been developed based on the idea of pedestrian's position estimation on an expected path [1]. This approach assumes a video of first-person vision that was pre-recorded along the path planned to go through with navigation. Reference points on the path are associated with images selected from the video. The previous approach selects images at constant interval. We propose an adaptive method of image selection that rejects images that are similar each other in the database. The comparison was made on real video for both of the constant interval and adaptive selection.

Keywords: Computer vision, Visual content-based indexing and retrieval, Augmented reality.

Index Terms: [Human computer interaction]: Mixed / augmented reality; [Computer vision]: Visual content-based indexing and retrieval

1 INTRODUCTION

Walking navigation system by GPS, has already been put into practice, but that is severely degraded and it is sometimes not available in indoors, undergrounds and urban canyons. In order to cope these situations, various sensors have been used or combined to locate pedestrian. One example of sensors used for indoor navigation is WLAN. Leppakoski [2] proposed a method using combined information with inertial sensor, indoor map and WLAN signals. Another example that was proposed by Ruiz [3] is a navigation system by coupling foot-mounted IMU and RFID. However, since these approaches are limited to indoor use, they depend on an infrastructure or could not be used in variety of situation from indoor to outdoor. The method of locating with cameras is also proposed [4]. Yet the cameras could improve the accuracy of pedestrian location estimation with other sensors, if the camera based location estimation can tell the position as highly reliable answer.

Kameda [1] proposed a method that estimates the pedestrian's position on the assumed path only using single camera (Figure 1). This method assumes reference points on the path associated with images of first-person vision. The images were selected from a video taken by someone who walked along the path in advance. The pedestrian's position is estimated by image retrieval between images selected from the video (image database) and a snapshot photographed by a pedestrian who wants to know the place. The location on the path is obtained by the location associated with the retrieved image in the database. We plan to conduct the precise camera registration with retrieved image and give 3D instruction for navigation as for our next development step.

Constant interval selection was proposed in the original approach [1]. It could be reasonable when the pre-recorded video was taken by someone walking at a constant speed. However, it is difficult to take such a video on real scenes of waiting for a traffic light, stopping to confirm safety, avoiding to collide with other pedestrians, etc. As a result, a number of very similar images from the pre-recorded video will be imported to the database for the same location where the camera was stacked for a while by the constant interval approach. It is a waste of memory and it will result in worse resolution of location estimation.

We propose a new method to adaptively change the interval of image selection in order to avoid successive similar images in the database. The comparative investigation through the experiment on real path demonstrates the efficiency of the proposed method.

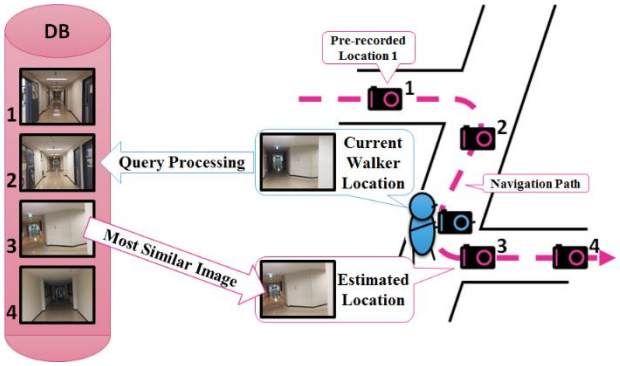

Figure 1: Location Estimation of the Pedestrian's Estimation by Image Retrieval Based on Pre-recorded Video of First-Person Vision.

2 LOCATION ESTIMATION

2.1 Similar Image Retrieval

We choose locality based operator SIFT [5] for image retrieval. SIFT keys have been calculated for all the reference images in the database and they are stored beforehand. On navigation, a single query image of pedestrian's first-person vision is taken and its SIFT keys are calculated immediately. Each key in the query image will be assigned to the most similar key in the database [6] if the keys are close enough. The similarity of the retrieved image is represented by the number of key-pairs found at the image. It can

†email: s1211104@u.tsukuba.ac.jp
‡email: {kameda, kitahara, ohta} @iit.tsukuba.ac.jp

Figure 2: Snapshots of the Pre-Recorded Video. (a) 200 frame. (b) 700 frame. (c) 1300 frame. (d) 1800 frame. (e) 2300 frame. (f) 2800 frame.

be said that the image that has the largest number of key-pairs was taken at the closest position from the pedestrian on the path. Therefore, the pedestrian's position is estimated based on the retrieval result.

2.2 Reference-Frame Selection

In the previous method, reference images are selected at constant interval in the pre-recorded video. In contrast, we propose a new adaptive method that selects images by changing interval. The interval is changed in order to avoid selection of consecutive similar images as reference images. To achieve this purpose, the video is scanned successively and the frames that are not similar to the previous one will be selected for reference images.

Suppose the pre-recorded video has N frames and each frame is indicated by $R_i (1 \le i \le \mathrm{N})$ sequentially. The number of SIFT keys extracted at R_i is denoted as M_i. A SIFT key at R_i is denoted by $k_j \cdot R_i (1 \le \mathrm{j} \le M_i)$ and its feature value described by 128-dimensional vector is represented by $k_j \cdot R_i \cdot descriptor$.

The similarity between R_i and R_k is given by

$$\mathrm{P} = \mathrm{CountSIFTPairs}(\mathrm{R_i}, \mathrm{R_k})$$

Where $\mathrm{CountSIFTPairs}(\mathrm{R_i}, \mathrm{R_k})$ counts the number of key-pairs between the query image R_i and a reference image R_k in the database.

On finding the most similar key for $k_j \cdot R_i$, false matching could be avoided by examining the similarity of two keys. The similarity of key-pairs are evaluated by

$$\mathrm{D} = \mathrm{DescriptorDistance}(\mathrm{k_j} \cdot \mathrm{R_i}, \mathrm{k_l} \cdot \mathrm{R_k})$$

This function calculates the distance in 128-dimentional space. The key-pairs that has longer distance than $\mathrm{D_{th}}$ will be removed. The procedure of our proposing adaptive reference image selection is shown in the algorithm 1. If the current investigating frame i looks different from the lastly selected reference frame k, the frame i will be treated as new reference frame. The frame similarity is controlled by the threshold parameter $\mathrm{P_{th}}$.The image set **I** will be used as the image database for a query from a user.

Algorithm 1 Adaptive reference image selection

Input: N is a set of images from pre-recorded video
Output: I is a set of selected images as references
1: **register** $\mathrm{R_1}$ **to I**
2: $\mathrm{k} = 1$
3: **for** $\mathrm{i} = 2$ **to** N **do**
4: **Find the most similar** $\mathrm{k_l} \cdot \mathrm{R_k}$ **for** $\mathrm{k_j} \cdot \mathrm{R_i}$ at $\mathrm{R_k}$
5: Remove Key-Pairs if
$\mathrm{DescriptorDistance}(\mathrm{k_j} \cdot \mathrm{R_i}, \mathrm{k_l} \cdot \mathrm{R_k}) > \mathrm{D_{th}}$
6: $\mathrm{P} = \mathrm{CountSIFTPairs}(\mathrm{R_i}, \mathrm{R_k})$
7: **if** $\mathrm{P} < (\mathrm{M_k} * \mathrm{P_{th}})$ **do**
8: $\mathrm{k} = \mathrm{i}$
9: register $\mathrm{R_k}$ to **I**
10: **end if**
11:**end for**

3 Comparative Experimental Results

We have conducted a comparative experiment on a real walking path of about two minutes that includes both outside scene and inside scene. The former part of the path is at outside (Figure 2 (a), (b), (c)) and the latter part is at inside, which includes stairs to the upper floor. The video for source of reference images was taken by 15fps at a size of 640x480 pixels. The number of images from pre-recorded video N is 3300. The video was taken by a pedestrian who stopped twice on the path, for about 600 frames at around 700th and 2500th frame. The number of extracted SIFT keys is shown in Figure 3. Since the path comes into the building at the 1840th frame, the number decreases beyond this point due to the less textures in the building. The number of reference images selected by the proposed method is 80 with the threshold value of $\mathrm{D_{th}} = 200$ and $\mathrm{P_{th}} = 0.1$.

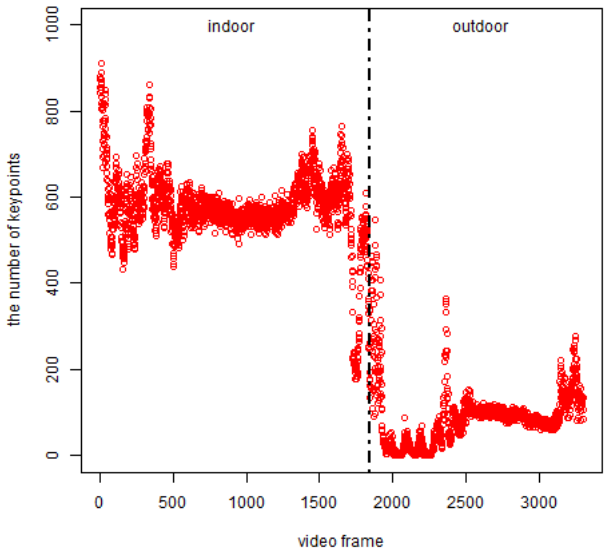

Figure 3: The Number of Extracted SIFT Key of Pre-Recorded Video.

As for queries, another video was taken just after the reference video was recorded in the same video format. The number of query images is 2100. The number of SIFT keys found in the query video is shown at Figure 4. The path comes into building at the 1220th frame.

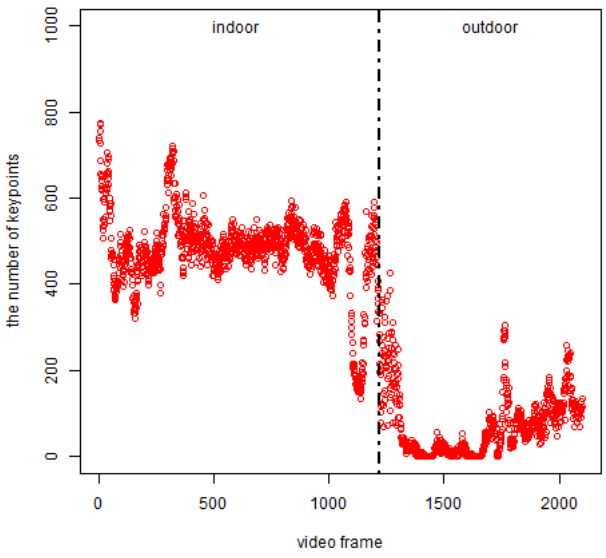

Figure 4: The Number of Extracted SIFT Key of Query Video.

The result of image retrieval for both the proposed adaptive selection method and the conventional constant interval method is shown in Figure 5. The horizontal axis indicates the No. of the reference frame in the database. The vertical axis corresponds to the original frame number in the pre-recorded video. As for the constant interval selection method, the interval is set to the same number of the adaptive method. In this experiment, images are selected at 40 frames interval. Note that the frames of stopping on the path are not selected in the adaptive method and more reference frames are assigned to the path where the camera was moving. This means that the adaptive method can have finer resolution on location estimation.

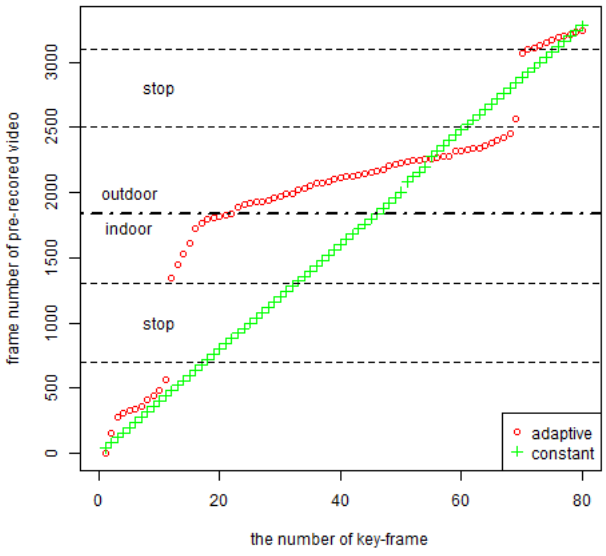

Figure 5: Selected Images as References.

The reason of the bent of the adaptive result at about 17th reference frame is the change of the scene, going into the inside from the outside. While the camera is at outside, objects are far and they have rich texture. It means the camera registration could be done well at relatively large interval. On the other hand, in the building, the floor, walls, and ceiling are relatively close to the camera and they have less textures. This means image registration may be easily fail if the interval is large. Therefore, our adaptive method automatically assigns more reference frames to the path inside the building.

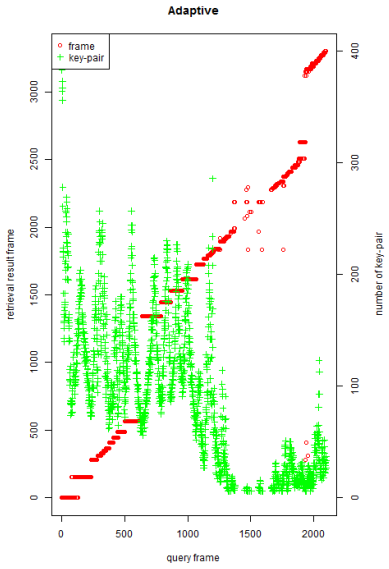

Figure 6: Image Retrieval of the Proposed Adaptive Method.

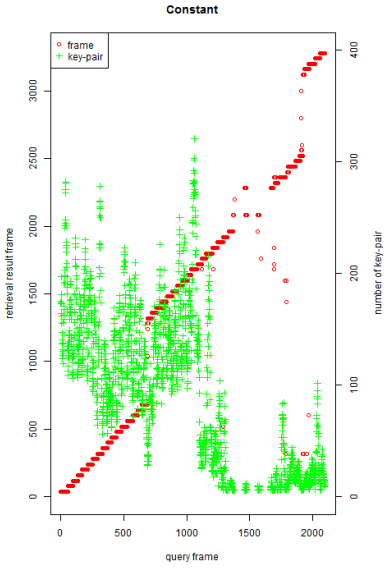

Figure 7: Image Retrieval of the Conventional Constant Interval Method.

Figure 6 shows the result of image retrieval of the proposed method. The query was issued for every frame of the video at 15 fps and the walking speed was same as the pre-recording except for the stopping periods. The horizontal axis means the number of the frame in the query video. The red dots indicates the number of the frame of the retrieved image. This means the red dots should be placed proportionally for normal walking part. Since the reference

images at outside part have rich texture and far objects in their images, they can cover relatively wider area compared with inside path. This change can be seen before and after about 700th frame of the query video. The green dots indicates the number of the key-pairs at each query.

The constant interval method can also shows very similar result in the experiment (Figure 7). Since this method assigns less reference images for the path inside the buildings, it may has more chances of losing the correct answers for that part. It failed in getting the correct reference frames around 1500th frame in the query.

4 Conclusion

We propose a new image selection method by changing the interval of the frames in the pre-recorded video adaptively so that successive similar images are avoided in the database. This results in assigning more frames to the path in motion. Our method also adaptively select the images so as to let each reference image corresponds to the wider area in the path, and it means more reference images could be assigned to critical part of the path.
We are going to develop the 3D navigation in AR fashion by conducting the 3D camera registration between the input query image and the retrieved image.

References

[1] Yoshinari Kameda and Yuichi Ohta. Image Retrieval of First Person Vision for Pedestrian Navigation in Urban Area. ICPR, pages 364-367, 2010.

[2] H.Leppakoski, J.Collin and J.Takala. Pedestrian Navigation Based on Inertial Sensors, Indoor Map, and WLAN Signals. Journal of Signal Processing Systems Volume 71, Issue 3, pages 287-296, 2013.

[3] A.R.J.Ruiz, F.S.Granja, J.C.Prieto Honorato and J.I.G. Rosas. Accurate Pedestrian Indoor Navigation by Tightly Coupling Foot-Mounted IMU and RFID Measurements, Instrumentation and Measurement, IEEE Transactions on Volume 61, Issue 1, pages 178-189, 2011.

[4] Ruotsalainen,L, Bancroft.J, Lachapelle.G, Kuusniemi.H and Ruizhi Chen. Effect of camera characteristics on the accuracy of a visual gyroscope for indoor pedestrian navigation. UPINLBS, pages 1-8, 2012.

[5] D.G.Lowe. Object recognition from local scale-invariant features. ICCV, pages 1150-1157, 1999.

[6] D.G.Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, Volume 60, Issue 2, pages 91-110, 2004.