A Visual-GPS Fusion Based Outdoor Augmented Reality Method

Junjie Wang *
School of Computer Science
Beijing Institute of Technology
Beijing, P.R.C
wangjunjie_ch@aliyun.com

Quanyu Wang
School of Computer Science
Beijing Institute of Technology
Beijing, P.R.C
wangquanyu@bit.edu.cn

Uzair Saeed
School of Computer Science
Beijing Institute of Technology
Beijing, P.R.C
uzairsaeed@bit.edu.cn

ABSTRACT

Virtual objects can be overlaid with real scenes through outdoor augmented reality technology, which bring about prominent experience. The existing outdoor augmented reality methods are usually limited in accuracy and scalability. To solve this problem, a novel method combining computer vision and Global Positioning System is proposed in this paper. The Geohash method is introduced to stimulate the retrieval of nearby locations. The vocabulary tree is built to recognize the current scene from the reference library. Faster R-CNN based object detection method is combined with AKAZE feature detection and image matching algorithm to realize the scene recognition and target tracking. The results show that our method can realize efficient and scalable outdoor augmented reality.

CCS CONCEPTS

• Human-centered computing~Visualization techniques

KEYWORDS

Outdoor Augmented Reality, outdoor localization, AKAZE, Geohash.

ACM Reference format:

Junjie Wang, Quanyu Wang, Uzair Saeed. 2018. A Visual-GPS Fusion Based Outdoor Augmented Reality Method. In *Proceedings of ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI2018). ACM, Hachioji, Japan , 4 pages.* https://doi.org/10.1145/3284398.3284414

1 Introduction

Augmented reality technology refers to the emplacement of computer rendered virtual image and the target object in the real world, thereby enhancing the user's perception and knowledge of the real world [Billinghurst et al. 2015]. AR technology has two main stages – Tracking and Registration [Billinghurst et al. 2015]. The traditional outdoor tracking technologies are usually based on the hardware devices, such as GPS, ultrasonic, electromagnetic waves and other hardware tracking technologies [Schall et al. 2009]. The hardware based outdoor tracking technology is high efficiency and invariant to light change and other external conditions. However, the accuracy of the method is poor [Feiner et al. 1997].To improve the accuracy of the outdoor AR tracking technology, vision-based method is introduced. Behringer et al combine the visual horizon silhouette segment with GPS information to realize the outdoor AR tracking for airplanes [Behringer et al. 1999]. A C/S framework vision based outdoor AR method is proposed in [Fu-Xiang et al. 2015]. Lazebnik et al, proposed a method, which integrates multi-sensors with edge-based tracker to realize the outdoor augmented reality [Reitmayr et al. 2007]. Rao, Jinmeng et al, combine deep learning and spatial relationship for geo-visualization to realize the mobile outdoor augmented reality [Rao et al. 2017]. These methods realize the outdoor augmented reality in different situations. The main idea of these methods can be divided into 3 categories: traditional vision based, visual-sensor based, and deep learning based [Luo et al. 2013]. The vision-based method mainly based on image matching. [Karami et al. 2017] The visual-sensor based methods combine computer vision with sensors; it can achieve accuracy recognitions, but is prone to severe motion blur, changes in illumination [Freeman 2016]. The deep learning method trains the model based on massive images and needs to train a new model when it comes to a new environment, which limits the scalability.

To solve these problems, the paper proposes a visual-GPS fusion method. The Geohash method is introduced to retrieval targets' information to solve the low efficiency of traditional distance calculation. As the recognition is realized by feature detection, it is not necessary to train a model based on massive images which leads to higher scalability. After the scene is recognized, we use a pre-trained model to detect the main part of the image and match it with the target. The results show that the new method has better performance than the traditional vision-based method and higher efficiency than direct distance calculation, it is able to realize scalable and efficient scene recognition for outdoor augmented reality.

2 Method

A client-server architecture is utilized to realize outdoor scene recognition in this paper, the overview of the framework is shown in Figure 1. The steps are as follows: (1) Collect scene data (location and reference images) in server terminal. (2) Search the

VRCAI '18, December 2–3, 2018, Hachioji, Japan
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6087-6/18/12…$15.00
https://doi.org/10.1145/3284398.3284414

target's data and builds a visual vocabulary tree based on the location of the user in server terminal. (3) Capture the current scene and submit it to the server after the main part detection and compression. (4) The feature vector of uploaded image is calculated based on the vocabulary tree and used to recognition. (5) Finally, the recognition result and the corresponding descriptors is returned to the client terminal.

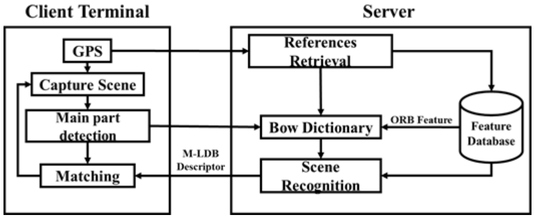


Figure 1: Framework of the Visual-GPS Based Method

2.1 Geohash Based Scene Information Retrieval

The location information is used to retrieve the outdoor scene information. Traditional location-based method usually calculates the distance by equation (1) [Rao et al. 2017]. Where Lat_a and Lon_a are the latitude and longitude of point A, Lat_b and Lon_b are the latitude and longitude of point B. R represents the radius of the earth. S is the distance between A and B. The method sorts different points by the distances between them, which is inefficient when the user is moving.

$$S = 2R\arcsin\sqrt{\left(\sin\frac{Lat_a - Lat_b}{2}\right)^2 + \cos Lat_a \times \cos Lat_b \times \left(\sin\frac{Lon_a - Lon_b}{2}\right)^2} \quad (1)$$

The Geohash algorithm is introduced to solve this problem. It is a latitude/longitude geocode system to convert Geographic coordinate system into a string using a base32 character map [Liu et al. 2014]. As shown in Figure 2, Geohash divides the earth into different grids. Every point in one grid shares a common Geohash string. Points in adjacent grids have similar Geohash string. Thus, we just need to search the scenes whose Geohash string is similar to current point.

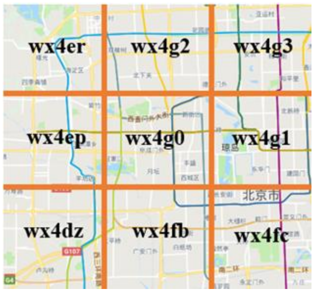


Figure 2: Geohash String of Different Grids on the Map

The accuracy can reach to 2cm when the maximum length of the string is 12 [Miller et al. 2010]. In this paper, the latitude and longitude of the target scenes are coded to a 9-length string where the accuracy is 5m. The nearest 6-7 scenes which are surround by the user's location will be retrieved when the sever requests the data.

2.2 Scene Recognition

To recognize the scene, the ORB feature of the target images are pre-stored. 6-7 scenes are set as the potential targets according to the location of the user. A visual vocabulary tree will be built based on the features of the target images by Bag of word Model [Tardos et al. 2012]. Then the captured scene will be recognized based on the vocabulary tree. The steps to recognize the scene are as follows:

1. Extract the ORB features of every target image.
2. Build the vocabulary tree based on the ORB features, and save the vocabulary into database.
3. Transform the ORB features to vectors based on the vocabulary tree for target images and captured scene.
4. Calculate and compare the distances between the target images' vectors and the scene's vector.

2.3 Main Part Detection and Image Matching

To reduce the background noises in outdoor environment, an object detection method is introduced to detect the main part of the image. A pre-trained model based on the Faster-RCNN [Ren et al. 2015] is used to detect the main part of images. The network is trained based on a large set of images. The location of these objects can be detected with the bounding boxes and will be extracted in later steps.

After the scene is recognized, the pre-stored M-LDB descriptors and the scene information will be returned to the client. In client terminal, the AKAZE algorithm is used to match the scene with the target. The wrong match results will be reduced by the RANSAC method. Moreover, a homography matrix, which indicates the pose relationship between the scene and the target, will be calculated in RANSAC process.

3 Result And Discussion

3.1 Geohash Based Scene Retrieval Result

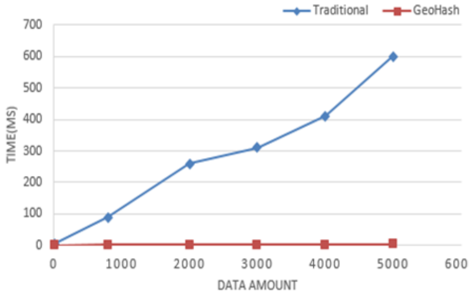


Figure 3: Runtime Comparison between Geohash Based and Traditional Method

The latitude, longitude and Geohash string are stored for every target scene. For Geohash based retrieval, we set the first several characters of the Geohash string as the search condition. The scope of the retrieval distance changes from 610 meters to 19 meters as the length of string ranges from 6 to 8 [Miller et al. 2010]. For traditional method, the distances between target scenes location and user's location are calculated and sorted. The results of two methods are shown in Figure 3.

It is showed that the Geohash based retrieval method is more efficient. When the amount of data increases from 0 to 5000, the runtime stabilizes at 2-4 ms, while the runtime of the traditional method increases linearly with the increase of data volume. The runtime of the traditional method reach to 600ms when the data volume is 5000. The results indicates that the Geohash based retrieval method is more efficient than the traditional method. Meanwhile, we can control the scope of the filter by the length of the Geohash string, while in traditional the threshold of the distance is fixed.

3.2 Result of Scene Recognition

The Geohash retrieval limit the amount of targets into 6 – 7. Then the vocabulary tree is built to recognize the scenes. To evaluate the performance of the method. We randomly select images from the INRIA dataset and oxford building dataset. These images are divided into 100 groups. In each group, there are 6 target scenes. For each scenes, there are 1-2 test images. The results are shown in Table I. It shows that the average time to build the vocabulary tree for each group is about 245ms. And the recognition time is about 5ms. The accuracy is 80.3% when the K equals 10 and the L equals 6 where K represents the node degree and L represents the height. The vocabulary file occupies 190KB for each group. The result indicates that the method can recognize the scene in real time and can satisfy the accuracy requirement.

Table 1: Performance of Recognition

Parameter	Accuracy	Train time(ms)	Recognize time(ms)
L=5 K=6	73.67%	247.12	5.36
L=5 K=9	75.63%	252.12	6.24
L=5 K=10	78.33%	261.76	6.67
L=6 K=6	74.7%	247.22	5.96
L=6 K=9	77.33%	258.22	6.24
L=6 K=10	80.3%	261.46	6.7

3.3 Result of Main Part Detection

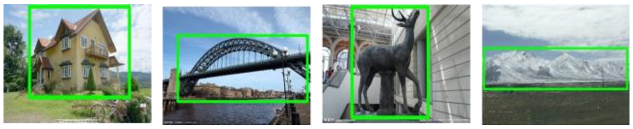


Figure 4: Main Part Detection of Selected Demonstrations

A pre-trained model is used to detect the main part of images. In this paper, we mainly test four outdoor scenes – buildings, status, bridges and mountains. Figure 4 shows the results of the main part detection of some demonstrations. The results show that the Faster-RCNN based detection method can detect the location of the main object in an image. We set 300 images as the test dataset for each classes to evaluate its mAP (Mean Average Precision). The results shows that the mAP can reach to 88%. The API of the main part detection is set on a remote sever; the average time to invoke the API is 219ms.

After the main part of the captured scene is extracted. Its feature points and the descriptors are computed by AKAZE algorithm, and the RANSAC is introduced to remove the wrong match points. To compare our method with original AKAZE method, two group of experiments are practiced, in each experiment several images are chosen to be matched. In the first group, the image is matched with the target without main part detection. In the second group, the main part of the image is extracted and is matched with the target scene. Figure 7 shows a set of selected demonstrations. Figure 5 (a) is the target, (b)-(c) are different scenes of the oriental pearl taken from different angles and lighting conditions.

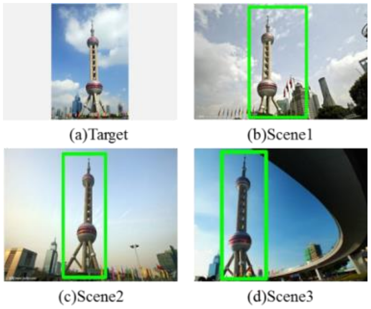


Figure 5: Target and Scene to be matched

Figure 6 shows the matching results of the first group. In Figure 6(a), few lines connect the target and the scene correctly. However in Figure 6 (b) and Figure 6 (c), most lines are wrongly connected. And the Figure 7 shows the marching results. It is obvious that the matching lines between the target and scenes are more accurate than the result shown in Figure 6. As the main part of the scene is extracted, the noises from the background are reduced. Thus, the accuracy of the image matching is improved. Meanwhile, the AKAZE algorithm needs to build an image pyramid and detect the feature points on it. As the size of the image is smaller, it will cost less time.

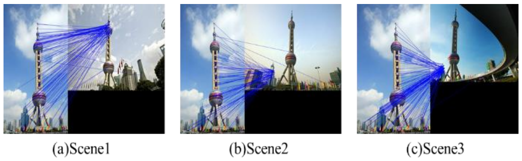


Figure 6: Matching Results of Original Scenes

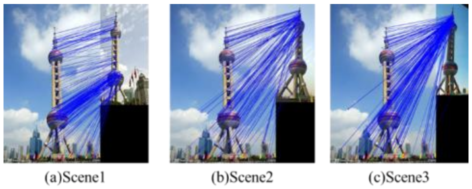
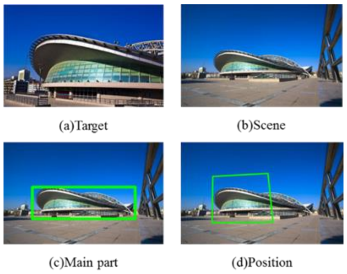


Figure 7: Matching Results of Extracted Scenes

Table II shows the comparisons between two methods from the runtime and accuracy. The inliers in table I means the amount of good matches between the target and the scene after the selection of RANSAC. It is found that the inliers in the second method are 2 to 5 times than that of the first method. Moreover, the runtime is 2 to 3 times more than the first method. Finally, the homography matrix will be calculated based on the matching result. It tells the pose relationship between the target and the scene. The exact position of the target in current frame is resolved by it. Figure 8 shows the result. It shows that the position of the target can be locate accurately in the scene.

Table 2: Matching Results Comparison between Original Scenes and Extracted Scenes

Name	Key points	inliers	Runtime(ms)	Ratio
Scene1	1796	126	812	0.07
Sub_scene1	896	229	300	0.26
Scene2	1170	89	774	0.07
Sub scene2	421	175	219	0.42
Scene 3	849	141	1184	0.17
Sub scene3	263	143	249	0.54



(a)Target (b)Scene

(c)Main part (d)Position

Figure 8: Position of the Target in the Scene

To realize the outdoor augmented reality, the location of the user used for retrieval the reference library to reduce the scope of the targets. The retrieval time cost about 4ms. The vocabulary tree is pre-trained before the recognition. It costs about 245ms. Once the vocabulary tree is built, it cost 5ms to recognize the current scene. The accuracy is about 80.3%. On the client terminal, the main part detection and the AKAZE algorithm is combined to realize image matching for target tracking. Apart from the training time, the total time is about 600ms. To add new targets, we just need to save the location and the reference image in our server. It shows that we are able to realize a real time and scalable outdoor augmented reality. However, as there are many similar outdoor scenes and that the content of the outdoor scenes are complex. The accuracy of the recognition remains to be improved.

4 Conclusion

To realize real time and scalable outdoor augmented reality. The Geohash based location retrieval method is introduced. The visual vocabulary tree is built to realize the scene recognition and main part detection method is combined with the AKAZE algorithm to realize the target tracking. The results show that the Geohash based method has a better and stable performance than the traditional method with the combination of the main part detection can improve the accuracy and performance of the AKAZE. The recognition time is about 5ms after the vocabulary is built. The total time of target tracking is about 600ms. In addition, the accuracy for outdoor recognition is about 80.3%, which can satisfy the requirement for outdoor augmented reality.

REFERENCES

Billinghurst, M., Clark, A., & Lee, G. 2015. A survey of augmented reality. Foundations and Trends in Human-Computer Interaction, 8(2-3), 73-272.

Schall, G., Wagner, D., Reitmayr, G., Taichmann, E., Wieser, M., & Schmalstieg, D., et al. 2009. Global Pose Estimation using Multi-Sensor Fusion for Outdoor Augmented Reality. IEEE International Symposium on Mixed and Augmented Reality (pp.153-162). IEEE.

Feiner, S., Macintyre, B., Hollerer, T., & Webster, A. 1997. A touring machine: prototyping 3D mobile augmented reality systems for exploring the urban environment. IEEE International Symposium on Wearable Computers (Vol.1, pp.74-81). IEEE Computer Society.

Behringer, R. 1999. Registration for Outdoor Augmented Reality Applications Using Computer Vision Techniques and Hybrid Sensors. IEEE Virtual Reality (pp.244). IEEE Computer Society.

Fu-Xiang, L. U., & Huang, J. 2015. Beyond bag of latent topics:spatial pyramid matching for scene category recognition. Frontiers of Information Technology & Electronic Engineering, 16(10), 817-828.

Reitmayr, G., & Drummond, T. 2007. Going out: robust model-based tracking for outdoor augmented reality. Ieee/acm International Symposium on Mixed and Augmented Reality (pp.109-118). IEEE.

Rao, J., Qiao, Y., Ren, F., Wang, J., & Du, Q. 2017. A mobile outdoor augmented reality method combining deep learning object detection and spatial relationships for geovisualization. Sensors, 17(9), 1951.

Luo, B., Wang, Y. T., Shen, H., Wu, Z. J., & Liu, Y. 2013. Overview of hybrid tracking in augmented reality. Acta Automatica Sinica, 39(8), 1185-1201.

Karami, E., Prasad, S., & Shehata, M. 2017. Image matching using sift, surf, brief and orb: performance comparison for distorted images.

Freeman, B. 2016. Computer vision overview. IEEE Expert, 6(4), 11-15.

Liu, J., Li, H., Gao, Y., Yu, H., & Jiang, D. 2014. A geohash-based index for spatial data management in distributed memory. International Conference on Geoinformatics (pp.1-4). IEEE.

Miller, F. P., Vandome, A. F., & Mcbrewster, J. 2010. Geohash. Alphascript Publishing.

Ren, S., He, K., Girshick, R., & Sun, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. International Conference on Neural Information Processing Systems (Vol.39, pp.91-99). MIT Press.

Zhong, X., Wang, W., & Wang, Q. 2017. An indoor AR registration technique based on iBeacons. IEEE International Conference on Information and Automation (pp.1093-1098). IEEE.

Tardos, J. D. 2012. Bags of Binary Words for Fast Place Recognition in Image Sequences. IEEE Press.Conference Name:ACM Woodstock conferenceConference Short Name:WOODSTOCK'18.