doi:10.3772/j.issn.1006-6748.2014.03.012

Outdoor guide system based on the mobile augmented reality technology[①]

Zhang Yunchao(张运超)[*], Chen Jing[②][**], Wang Yongtian[*][**], Xu Zhiwei[**]
([*]School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, P. R. China)
([**]School of Optics and Electronics, Beijing Institute of Technology, Beijing 100081, P. R. China)

Abstract

This paper proposes an outdoor guide system using vision-based augmented reality (AR) on mobile devices. Augmented reality provides a virtual-real fusion display interface for outdoor guide. Vision-based methods are more accurate than GPS or other hardware-based methods. However, vision-based methods require more resources and relatively strong computing power of mobile devices. A C/S framework for vision based augmented reality system is introduced in this paper. In a server, a vocabulary tree is used for location recognition. In a mobile device, BRISK feature is combined with optical flow methods to track the offline keyframe. The system is tested on UKbench datasets and in real environment. Experimental results show that the proposed vision-based augmented reality system works well and yields relatively high recognition rate and that the mobile device achieves real-time recognition performance.

Key words: mobile augmented reality, location recognition, vocabulary tree, optical flow, tracking and registration

0 Introduction

Augmented reality (AR) is an important research direction in the computer vision field. Computer-generated virtual objects can be projected into the real environment seamlessly. The technology of mixed virtual and actual reality is widely applied in guide services, such as location-based service. SPRXmobile developed the first mobile augmented reality browser Layar in the world in 2009. The browser offers various items of surroundings based on augmented views. Texts or image information associated with geographical position are laid over the camera view. Nokia City Lens is also a popular augmented reality browser which can provide dynamic information about users' surroundings such as hotels, attractions in the forms of virtual objects overlaid above buildings.

Most of traditional AR browsers such as Layar and City Lens are based on hardware sensors to locate user's positions. Those hardware sensors include GPS, electronic compass and the accelerometer. However, built-in sensors on the cell phone such as GPS usually have low sensitivity and others may have cumulative errors. By contrast, the vision-based location recognition method can achieve sub-pixel accuracy and a more authentic experience for users. Location recognition and mobile tracking are the key technologies for an outdoor AR browser. Now, outdoor location recognition remains a challenge for many reasons. Occlusion, illumination change, repetitive structure of outdoor buildings are factors affecting the recognition accuracy. Mobile tracking with six degrees of freedom (6DOF) has to face the complex changes of scale and lighting conditions. It must give absolute measurements with respect to a given coordinate system, which is very robust and runs in real-time. The limited memory and computing power of the mobile devices greatly influence the robustness and real-time performance of vision-based AR browser.

In this paper, a C/S framework for mobile AR browser is proposed, which can be used in city-scale outdoor guide system. The main contributions of this work are summarized as follows:

(1) A vocabulary tree algorithm[1] is employed with an initial clustering center selection method for location recognition. The selection of initial clustering center can speed up the training of the vocabulary tree and improve the clustering effect of the vocabulary tree.

[①] Supported by the National Science and Technology Major Project (No. 2012ZX03002004) and the National High Technology Research and Development Programme of China (No. 2013AA013802).

[②] To whom correspondence should be addressed. E-mail: chen74jing29@bit.edu.cn

Received on June 18, 2013

(2) BRISK[2] feature is combined with optical flow to achieve robust and real-time tracking of mobile devices.

(3) An image and feature compression strategy is applied for the network transmission without affecting the precision of recognition rate.

1 Related work

Location recognition is closely related to image retrieval problems. For large-scale image retrieval, the commonly adopted scheme is the machine learning method, which utilizes classifiers to do supervised learning of image features and then transfers image retrieval to a feature classification problem. Recently, researchers have proposed support vector machines (SVM), K nearest neighbor (KNN), principal component analysis (PCA), random Ferns and other scene recognition algorithms[3-7], most of which have low processing speed and are unable to meet the real-time requirements of the AR system. The random Ferns method has relatively high recognition rate and processing speed. However, the random Ferns method requires a lot of memories during operation and cannot be applied in a city scale location recognition. David Nister applied the vocabulary tree algorithm for scene recognition[1] and achieved high recognition rate. The closest work to ours is probably carried out by Schindler[8] and Baatz[9]. They both adopted Nister's vocabulary tree for location recognition. The traditional vocabulary tree including Nister's method is based on a randomly selected initial feature clustering center. For large-scale image feature clustering, it will be time consuming. The clustering algorithm often clusters slowly and does not even converge due to random clustering center. Although, Schindler[8] improved the scoring method of visual words, and Baatz[9] introduced image preprocessing and re-ranking method, the traditional clustering method affects the recognition rate. In our work, we propose a novel selection method for initial feature clustering center, which can get good clustering center quickly and avoid unnecessary iterations.

Real-time and robust tracking methods also have received a wide range of attentions, especially methods applied on mobile devices. Wagner[10] described modified SIFT[11] and Ferns[6] approaches and created the first real-time 6DOF natural feature tracking system running on mobile phones. Klein and Murray[12] introduced a keyframe-based SLAM[13] system on a camera phone. They proposed a series of adaptations to the Parallel Tracking and mapping system to mitigate the impact of the device's imaging deficiencies. These above mentioned methods can be only applied in a small work area. Leutenegger proposed the BRISK[2] method for keypoint detection, description and matching which have adaptive, high quality performance at a dramatically lower computational cost. The key to speed lies in the application of a novel scale-space FAST-based detector[14] in combination with the assembly of a bit-string descriptor[15] from intensity comparisons retrieved by dedicated sampling of each keypoint neighbourhood. Compared with SIFT and SURF[16], BRISK has real-time performance. In our system, BRISK feature is employed to initialize the mobile tracking and optical flow is used to speed up subsequent computing.

2 Methods

In this paper, a client-server architecture is utilized to realize location recognition and mobile tracking, and an overview of the framework is given in Fig. 1, which contains three modules, i. e. location recognition, mobile tracking and network transmission:

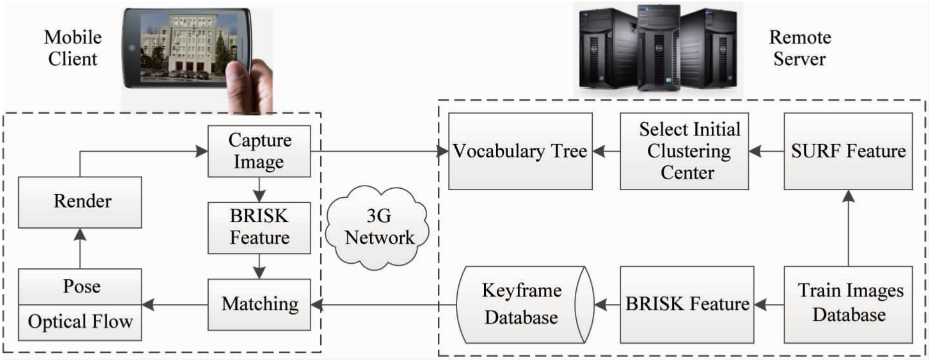


Fig. 1 Framework of the mobile augmented reality system

Location recognition. In the offline stage, SURF features are extracted from the training images stored on the remote server database. These features are then quantized into a vocabulary of visual words with a hierarchical k-means clustering scheme. A selection method of initial clustering centers is applied to the original k-means algorithm. In the online stage, query images from mobile devices are fast indexed by the above tree structured vector quantization.

Mobile tracking. BRISK features are offline extracted to establish the keyframe database from the training images stored on the remote server. Two types of feature tracking is combined to realize real-time performance. BRISK features are detected and tracked frame-to-frame by computing optical flow. The point correspondences between features in the captured frame and the offline-built keyframe can be used for estimating a camera pose.

Network Transmission. A JPEG-compressed query image is transmitted from a mobile device to a remote server and then queries a database hosted on the server. Then, compressed keyframe features are transmitted to the mobile device.

2.1 Location recognition

The vocabulary tree is widely used for the recognition of large scale images. It has relatively high recognition rate and speed which is suitable for our outdoor location recognition. Image features are offline extracted and clustered to build a vocabulary tree. The scores of visual words are computed with the TF-IDF model according to Eq. (1). $w_{i,j}$ is the weight of image j on visual word i. $m_{i,j}$ is the count of visual word i on image j. N is the total number of train images and n_i is the count of images which includes visual word i.

$$w_{i,j} = m_{i,j} \times \lg \frac{N}{n_i} \qquad (1)$$

A similar quantification strategy is used for query images online. The query process is provided as

$$S(\boldsymbol{d}_j, \boldsymbol{q}) = \| \boldsymbol{d}_j - \boldsymbol{q} \|_p \qquad (2)$$

Generating clusters for such a large quantity of data presents challenges to traditionally used algorithms. For large-scale location recognition, clustering time of mass images is very long as the initial clustering center is randomly generated. In order to obtain ideal clustering results, we designed the selection methods of initial clustering center based on the traditional vocabulary tree as follows:

Step 1. Select only one clustering center randomly.

Step 2. Compute $D(x,K)$, the minimum distance between other point x and existing cluster centres.

Step 3. Add one new point as a clustering center. Each point x is chosen with the probability proportional to $D^2(x,K)$, as shown in

$$P(x \in K) \sim \min D^2(x,K) \qquad (3)$$

Step 4. Repeat Steps 2 and 3 until k centres have been chosen.

Distinct initial clustering centers could be found by this method, which reduces the number of iterations greatly and achieves good clustering performance.

2.2 Mobile tracking and registration

In most cases, the observer is far away from the target buildings and the building surface can be approximated as a plane. With keypoint matching, we can establish the homography between current frame and keyframe according to

$$sm_c = H_k^c m_k \qquad (4)$$

where m_c is a feature point in the captured frame and $m_k$is the matching point in the keyframe. s is the scale factor . The point correspondences are built as the tracking initialization step through BRISK feature matching . Real-time performance is achieved by calculating the optical flow between successive frames. The optical flow measurements are refined with tracking to avoid the drift introduced by frame-to-frame feature matching. An initial H_k^c is obtained with four matching points. We select H_k^c with the minimal back-projection error d according to Eq. (4) as our homography.

The solution to Eq. (5) can provide a reasonable estimate of the camera pose, yet typically leads to the jitter problem, particularly noticeable when the camera is fully or nearly stationary.

$$d = \| \boldsymbol{m_c} - \boldsymbol{H_k^c m_k} \| < q_{pro} \qquad (5)$$

In order to stabilize the solution, we use the pose estimation results as initial data and perform some optimizations, such as the L-M optimization method.

2.2.1 BRISK feature matching

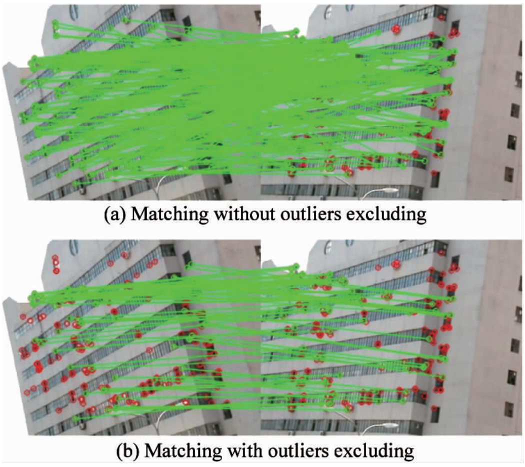
BRISK feature is a kind of binary feature based on AGAST feature detection and a BRIEF feature description. BRIEF can achieve scale invariant by building an image pyramid. For BRIEF feature local binary sampling is applied to generate binary descriptors, as shown in Eq. (6). $I(P_j^a, \sigma_j)$ is the gray value of the sampling point by Gaussian smoothing.

$$b = \begin{cases} 1, & I(P_j^a, \sigma_j) > I(P_i^a, \sigma_i) \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

Hamming distance is used to measure the difference among the feature descriptors. Calonder[16] has indicated the effectiveness of Hamming distance for classifying pairs of points. The distribution of the distance for non-matching points is Gaussian distribution and is centred around half of the descriptor dimension. The number of matching feature points can be controlled by limiting the Hamming distance, as shown in

$$d(\boldsymbol{m}, \boldsymbol{n}) = \sum_{i=1}^{k} m_i \oplus n_i < q \qquad (7)$$

PROSAC[17] is used to exclude outliers for binary descriptors after sorting the descriptors, as shown in Fig. 2.


(a) Matching without outliers excluding

(b) Matching with outliers excluding

Fig. 2 Excluding outliers with PROSAC

2.2.2 Optical flow tracking

Feature tracking can be considered a small-baseline tracking problem as the transformation between two consecutive image frames can be modelled using the translational model. Widely used feature tracking methods such as tracking by detection consume a lot of time for mobile devices and may affect the real-time performance. Computing the optical flow for the feature points provides us a real-time algorithm of frame-to-frame feature tracking, as expressed in Eq. (8). The motion of pixels can be estimated by the spatial derivative (I_x, I_y) and the temporal derivative I_t with a fixed size window.

$$I_x u + I_y v + I_t = 0 \tag{8}$$

With optical flow, we can speed up the computation of the camera pose, as shown in Fig. 3.

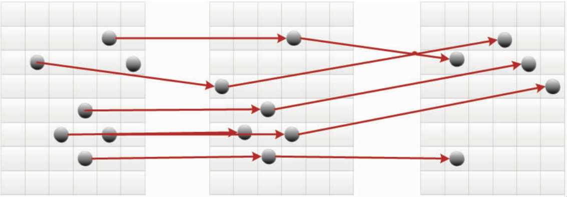


Fig. 3 Optical flow speed up

2.3 Network transmission

Large-scale image recognition is very strong in demanding requirements for network transmission. System fluency can be improved through minimizing the burden of transmission.

In this paper, we design the two-way transmission between the mobile phone client and the remote server as shown in Fig. 4.

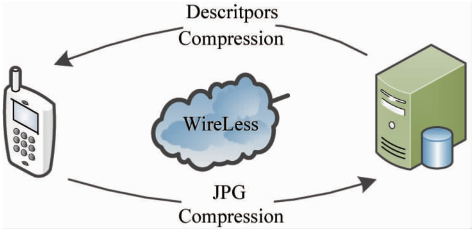


Fig. 4 Network transmission

Query image is transmitted in the form of JPG compression from a mobile device to the remote server. At the same time, BRISK features are transmitted in the form of Zip compression from the server to a mobile device. Most smart phones support the output of JPG image with different compression rates and are simple to be realized. The size of image can be quickly reduced to 1/3 of original size without affecting image quality. This kind of compression almost does not effect on the image feature detection. We will test it in our following experiments.

3 Experimental results

The system is tested on UKbench datasets and real environments, as shown in Fig. 5. The server is a high performance workstation with Intel Core i7 Processor and the mobile device is HTC 1.2GHz smart phone. 3G network is used for transmission.

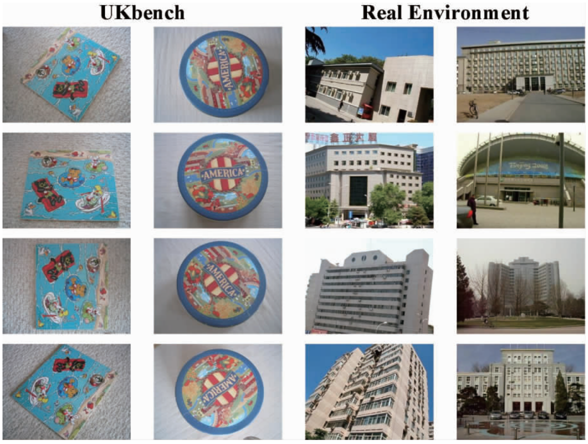


Fig. 5 UKbench datasets and real environment

3.1 Recognition performance

UKbench contains 2550 scenes and each scene contains 4 different images, including different rotation and lighting conditions as shown in Fig. 5. One of the 4 images is set as the training image and others as the query images.

The clustering time of the selection method of initial clustering center is compared with random initial clustering center used in Ref. [1]. Clustering time is shown in Table 1.

Table 1 Cluster time (s)

Initial Center	K = 8, L = 4	K = 8, L = 6	K = 10, L = 4	K = 10, L = 6
Nister	2513.21	2913.41	3818.56	6179.13
Our	1336.16	1357.40	2028.83	2810.76

The recognition accuracy is compared with that obtained by Schindler[7] and Baatz[8]. In order to accurately verify the clustering method, the same re-ranking method is employed to Baatz's and the recognition rate of top-10 mAP(%) is counted, as shown in Table 2. Experiment results show that our selection algorithm of initial clustering center can save the clustering time and enhance the recognition rate.

Table 2 Recognition rate (%)

Method	K = 8, L = 4	K = 8, L = 6	K = 10, L = 4	K = 10, L = 6
Schindler	73.13	80.26	78.53	81.79
Baatz	76.16	84.13	82.64	85.13
Our	79.06	86.55	84.58	88.96

Our recognition efficiency is also tested online. The recognition can be divided into four parts: feature detection of query image, bag of feature vector quantization, searching the inverted file of vocabulary tree and re-rank of retrieval short-lists. Time cost of each part is provided in Table 3.

Table 3 Online Recognition (ms)

Query	BOW	Search	Rerank	Total
103.1	0.7	6.3	65.1	175.2

3.2 Mobile tracking performance

The SURF feature is widely used in real-time tracking. The BIRSK-based tracking strategy is compared with the SURF-based tracking strategy and the time cost of different methods is counted, as shown in Fig. 6.

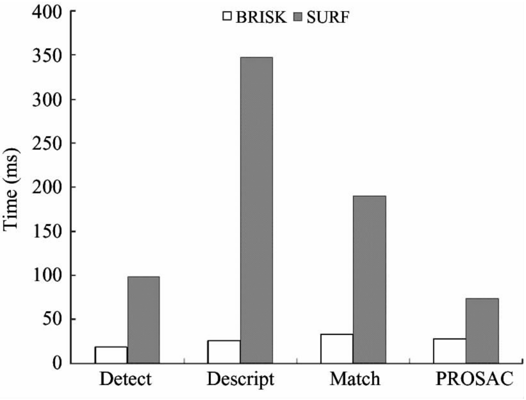


Fig. 6 Comparison of time cost between different features

The BRISK feature is faster than SURF and its total time cost of computing homography is about 100ms per frame. For most applications, ten frames per second cannot satisfy users. The optical flow method is used for real time tracking purpose.

We set the BRISK feature detection and matching as the initial step of mobile tracking and introduced optical flow to track the following frames. As the drift of tracking points and outliers excluding, the tracking points will become less and less with time, as shown in Fig. 7, where the number of tracking points and time cost of 80 frames are recorded.

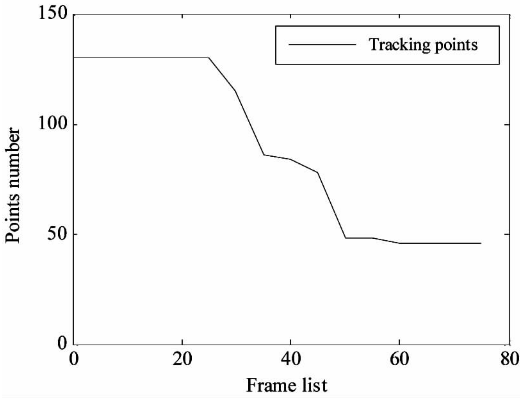


Fig. 7 Optical tracking points changes per frame

While the tracking points are too low, a re-initialization thread is conducted, and the time cost of optical flow is lower than that of initial proceeding. We count the time cost of optical flow computing and PROSAC per frame in Fig. 8. In most cases, frame rate can reach 25 or more.

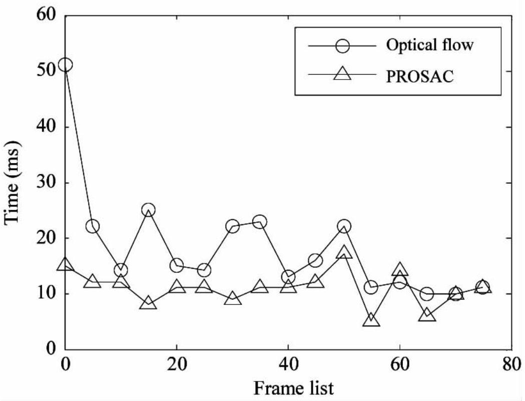


Fig. 8 Optical flow time cost changes per frame

3.3 JPG compression

The influence of JPG compression on recognition rate is tested. As shown in Fig. 9, the file size of tested image quickly decreases with JPG compression rate. By contrast, the recognition rate of vocabulary tree decreases slowly. File size will be reduced to 1/3 of original size when the compression parameter is set to 50. The recognition rate is almost constant, as shown in Fig. 10. Experiment results show that, JPG compression does not show much adverse effect within some compression rate. We can transmit JPEG-compression images without affecting recognition rate.

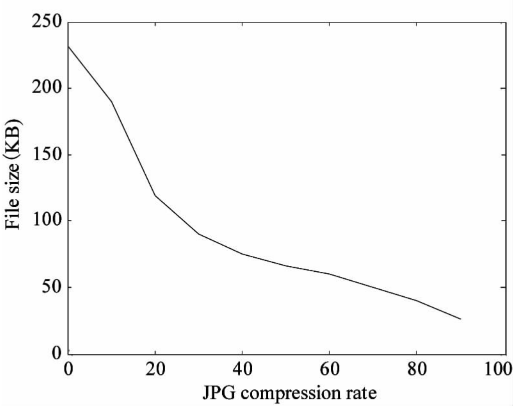


Fig. 9 JPG compression Effect on file size

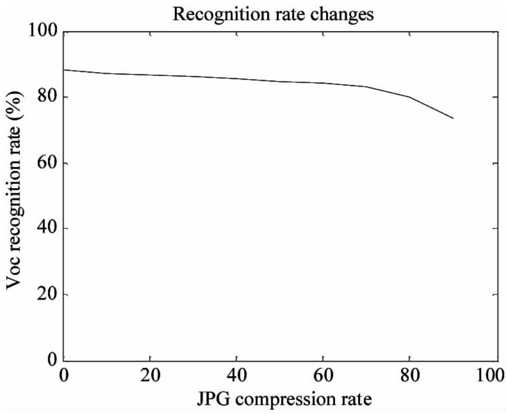


Fig. 10 JPG compression Effect on recognition

3.4 Applied system

The system is also tested in the outdoor environment, as shown in Fig. 11. Training images are collected from the Internet and real environment. Vocabulary tree is built and augmented data including virtual signs and 3D objects are stored in the database. 3G network is used for transmission. Real-time performance of the whole system can be seen in Table 4. Initialization including online recognition requires about 300 ms. For large-scale city guide, the speed can be accepted. Initial step only needs to be performed once before online tracking. During the online tracking stage, optical flow computing and 3D rendering may take less than 50ms in total. With more than 20 frames per second, we can realize a smooth tracking interface.

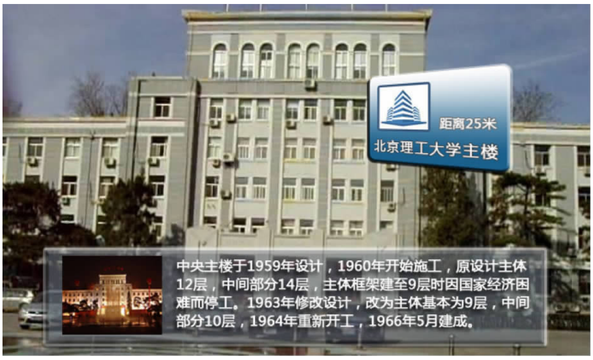

距离25米
北京理工大学主楼
中央主楼于1959年设计，1960年开始施工，原设计主体12层，中间部分14层，主体框架建至9层时因国家经济困难而停工。1963年修改设计，改为主体基本为9层，中间部分10层，1964年重新开工，1966年5月建成。

Fig. 11 Outdoor guide samples in the school

Table 4 Realtime performance (ms)

3G Transmission	Online Recognition	Initial Tracking	Optical Flow	3D Rendering
6.7	168.3	96.3	32.3	15.1

4 Conclusions

Experiment results show that the C/S system works well under complex outdoor environment. With the selection of initial clustering centers, vocabulary tree can be quickly built and good recognition rate can be reached. BRISK feature and optical flow running in parallel allow robust and real-time tracking. However, many problems still exist. The computing power of smart phones is still lower than that of PC. The smartphones have its advantage of the built-in sensors. We can narrow the scope of image retrieval with GPS and realize more accurate tracking with inertial sensors. Combining vision-based methods with hardware, we can realize robust and real-time augmented reality.

References

[1] Nister D, Stewenius H. Scalable recognition with a vocabulary tree. In: Proceedings of the 2006 IEEE International Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2161-2168

[2] Leutenegger S, Chli M, Siegwart R. BRISK: Binary robust invariant scalable keypoints. In: Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, 2011. 2548-2555

[3] Schmid C, Lazebnik S, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE International Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2169-2178

[4] Berg A, Zhang H, Maire M, et al. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proceedings of the 2006 IEEE International Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2126-2136

[5] Zisserman A, Bosch A, Munoz X. Representing shape with a spatial pyramid kernel. In: Proceedings of the 2007 ACM International Conference on Image and Video Retrieval, Amsterdam, Netherlands, 2007. 401-408

[6] Calonder M, Ozuysal M, Lepetit V, et al. Fast keypoint recognition using random ferns. *IEEE Trans PAMI*, 2010, 32(3): 448-461

[7] Wang Y S, Gao W. Pornographic image detection based on visual words and semantic projection. *Chinese High Technology Letters*, 2009, 19(10):1041-1047 (in Chinese)

[8] Schindler G, Brown M, Szeliski R. City-scale location recognition. In: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007. 1-7

[9] Baatz G, Koeser K, Chen D, et al. Handling urban location recognition as a 2D homothetic problem. In: Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 2010. 266-279

[10] Wagner D, Reitmayr G, Mulloni A, et al. Pose tracking from natural feature on mobile phones. In: Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality, Orlando, Cambridge, UK, 2008. 125-134

[11] Lowe D. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2):91-110

[12] Klein G, Murray D. Parallel tracking and mapping on a camera phone. In: Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality, Orlando, USA, 2009. 83-86

[13] Smith R, Self M, Cheeseman P. Estimating Uncertain Spatial Relationships in Robotics Autonomous Robot Vehicles. Autonomous Robot Vehicles. Berlin: Springer, 1990. 167-193

[14] Mair E, Hager G, Burschka D, et al. Adaptive and generic corner detection based on the accelerated segment test. In: Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 2010. 183-196

[15] Calonder M, Lepetit V, Strecha C, et al. BRIEF: binary robust independent elementary features. In: Proceedings of the 11th European Conference on Computer Vision, Heraklion, Greece, 2010. 778-792

[16] Bay H, Ess A, Tuytelaars T, et al. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 2008, 110(3):346-359

[17] Chum O, Matas J. Matching with PROSAC-progressivesample consensus. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005. 220-226

Zhang Yunchao, born in 1987. He received his B. S. degree from Shandong University in 2010. He is currently a Ph. D candidate in Beijing Engineering Research Center for Mixed Reality and Novel Display Technology, School of Optics and Electronics, Beijing Institute of Technology. His research interest include augmented reality and virtual reality.