International Journal of Future Generation Communication and Networking
Vol. 9, No. 6 (2016), pp. 165-178
http://dx.doi.org/10.14257/ijfgcn.2016.9.6.16

The Mobile Visual Search Guiding System Based on SIFT

Gongwen Xu[1,*], Xiaomei Li[2], Honglan Zhou[3], Jian Lei[4] and Zhijun Zhang[1]

[1]Shandong Jianzhu University, School of Computer Science and Technology, Jinan, China 250101
[2]Cancer Center of the Second Hospital, Shandong University, Jinan, China 250013
[3] Dongying Special Equipment Inspection Institute, Dongying, China 257091
[4] Shandong Institute of Standardization, Dongying Branch, Dongying, China 257091
[*]xugongwen@163.com

Abstract

In order to provide personal guiding service to visitors in the Picture Gallery, a mobile visual search guiding system based on SIFT(Scale Invariant Feature Transform) is proposed in this paper. The visitors can achieve the paintings' information using the camera function of the mobile phones when they enjoy the famous paintings in the Picture Gallery. The mobile search visual system proposed in this paper can identify the paintings in the server side so there are no extra performance requirements for the mobile phones. Firstly, the mobile visual search system was introduced. Then the system was designed and realized. The classification training was based on SVM (Support Vector Machine). The SIFT feature extraction algorithm was used to enhance the recognition rate. After the image retrieval, the geometric consistency check was employed to choose the best matching image. Finally, the experimental results verified the feasibility and practicability of this system. Compared with the present guiding systems, this system can communicate with the visitors conveniently and offer abundant and all-around information to the visitors. In addition, using this guiding system, any auxiliary devices are not needed to be installed in the Picture Gallery.

Keywords: *mobile visual search, picture gallery, guiding system, SIFT*

1. Introduction

Nowadays, information technology is applied in many daily fields to make people's life more convenient. At the same time, the informationization construction of the Picture Gallery is the general trend. The traditional narrator guiding method cannot meet with the visitors' need well, and some kinds of digital guiding systems emerged with the development of the times.

Now there mainly are two kinds of guiding systems [1]. The first one is the digital touch-tone design, whose client side is a handset like mobile phone. The visitors can select the aiming paintings by touching the keyboard to play the contents that were stored in it in advance. This design mode is the most-used one which needn't install serving devices in the Picture Gallery. The second one is the auto-sensing design. In this design, the automatic recognition function is added on the digital touch-tone mode, which resolves the troubles of the hand inputting information. At the same time, the Bluetooth emitters, RFIP tags and other auxiliary devices are needed to be installed in Gallery. Not all the Galleries permit the install of these devices, so this design is limited to be used and popularized [2].

The above systems cannot provide enough information and cannot interact with the visitors conveniently; besides, the visitors need to hire the handheld devices so they cannot serve the visitors well.

In this paper, aiming at the above problems, a mobile visual search system used in Picture Gallery guidance was proposed. The system based on SIFT can identify the showing paintings using mobile terminals. The remainder of this paper is organized as follows: Section 2 introduces the mobile visual search system. The design of the system is proposed in section 3. In section 4, the realization based on SIFT was described. The design of the mobile user interface is introduced in section 5. In section 6, the experiments and performance evaluation were reported. Finally, the conclusions and future works were given in section 7.

2. The Mobile Visual Search System

The mobile visual search means that the querying image of the object is gained by the mobile terminal in the real world. The corresponding information about the object is sent out via wireless network. So this retrieval method is based on the mobile terminal and wireless network [3].

The mobile visual search system gains the wide attention and researching effort in the academic field and industrial field. But in the real application, the performance of the visual search system is unstable and don't do well. Each unfavorable research result will harm user experiment. On the server side, the retrieval performance is the most important thing to deal with. Another important thing is the challenge coming from the limited wireless bandwidth. In the low bandwidth wireless network, upload of large amount of images can add the transmission delay and affect the user experiment. In this paper, the experiment was carried out in the WiFi network, which is suitable for the Picture Gallery environment.

The mobile visual search system in this paper uses the mobile phone camera to gain the paintings and search the interesting things. In this method, the objects are recognized by computing visual methods and the aiming objects are connected with the corresponding virtual digital information. The user need not input any keywords to retrieve the interesting information, so it is a humanized information retrieval method.

There are mainly two kinds of mobile visual search frames nowadays, terminal recognition and server recognition.

The terminal recognition frame is not widely used. The most representational one is the PhoneGuide system designed by Bauhaus University which can recognize some displaying items[4]. The Bluetooth signal identification is used to locate the visitor's position in museum to reduce the number of the identifying objects and enhance the speed and precision of the recognition. The advantage of this frame is that the mobile terminal can recognize the objects, and the network communication is not necessary. The drawback of this frame is that only few objects can be recognized one time and the performance of the mobile phone should be higher. Otherwise, the extra locating devices such as Bluetooth emitters are required in Picture Gallery.

The server recognition frame can recognize huge amounts of objects in the server side. The Google Goggles is the representational one which can recognize words, books, landmarks and artworks. This system has little performance requirement for the mobile phone. But the recognition speed is lower due to the communication delay and the large transporting data. The time of the recognition is more than 3 seconds, so the user experience is not good.

According to the questions in the guiding system, the mobile visual search system was modified in this paper. In this frame, the huge amounts of the objects on show can be identified on the server side, and the fast recognition method was adopted to improve the recognition performance.

3. Design of the System

3.1. The Structure of the System

The Client/Server construction is used to support more users and transport huge amounts of paintings information. The structure is shown in figure 1, composed by intelligent mobile, server, paintings information database. Mobile phone connects to server via 3G or WiFi network. The paintings information database stores the words, pictures, audio and video about the showing paintings to serve visitors.

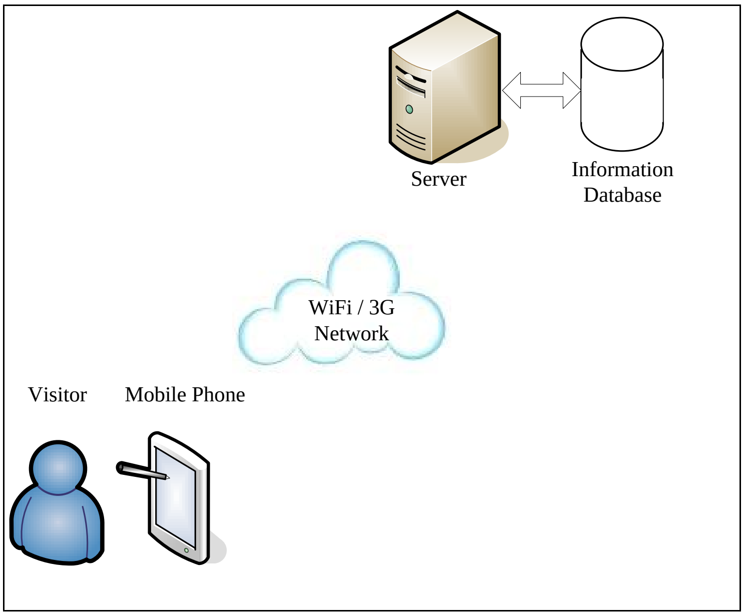


Figure 1. The Structure of the Guiding System

Visitors can take photos of the interesting paintings on show. This system can recognize the paintings automatically, and retrieve the database on the basis of the recognizing results. The words, pictures, audio and video are offered to visitors according to their requirement. Otherwise, this system has considered the different performance, function and processing ability of the different mobile phones.

Whether the mobile phone has high or low performance, the image recognition is carried out in the server side. When visitors take a photo on the interesting painting with mobile phone, the mobile application will connect the server automatically and send the painting to the server. The server recognizes the image, and then outputs the identifier of the painting. The corresponding information of this painting will be retrieved from the database and displayed on the visitors' mobile screen. The user experience is comfortable.

3.2. The Image Feature Extraction

The image feature extraction module mainly extracts the low level visual feature of the image. In this paper the image color feature and textual feature are extracted using image content describing technology. The color feature is described by color histogram and color matrix [5]. The textual feature is described by LBP (Local Binary Pattern) operators [6-7]. The flow chart of the process is shown in Figure 2.

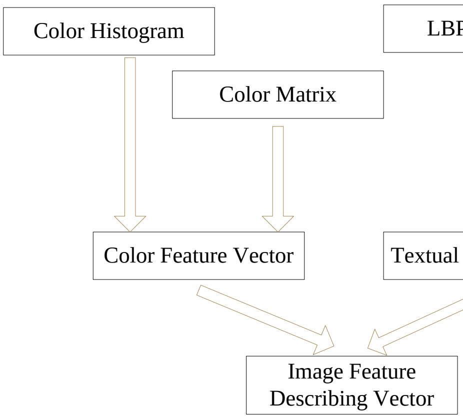


Figure 2. The Flow Chart of Feature Extraction

All the feature describing vectors are combined into image low level visual descriptors.

3.3. The Classification Training Module

The task of this module is to bring about the SVM (Support Vector Machine) classification training [8]. Firstly the training opposite samples and negative samples are selected, then feature vectors of all the training sample images are calculated using feature exaction module, at last the SVM classification is trained. The open source LibSVM is used to train SVM classification in this paper [9]. The flow chart of the process is shown in figure 3.

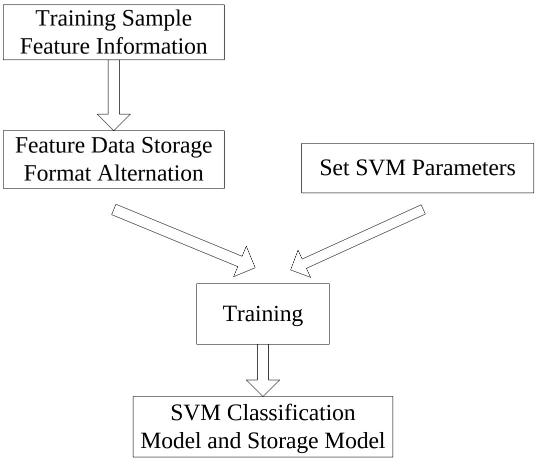

Training Sample Feature Information
Feature Data Storage Format Alternation
Set SVM Parameters
Training
SVM Classification Model and Storage Model

Figure 3. The Flow Chart of Classification Training

3.4. The Image Retrieval Module

The task of this module is completed in the server side and the image retrieval frame is based on content. The flow chart of this process is shown in Figure 4.

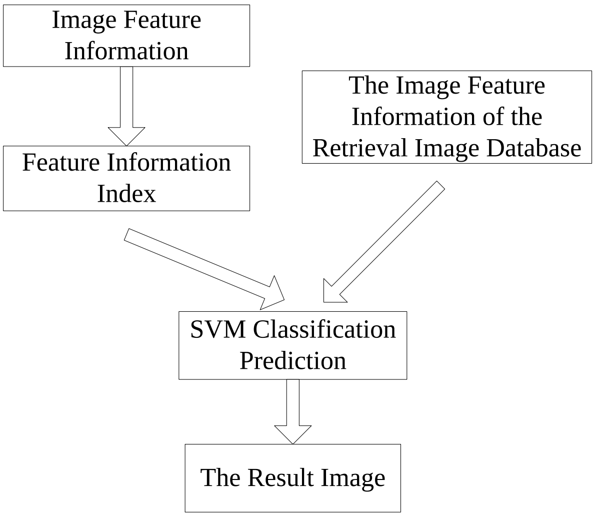


Image Feature Information
Feature Information Index
The Image Feature Information of the Retrieval Image Database
SVM Classification Prediction
The Result Image

Figure 4. The Flow Chart of Image Retrieval

It can be seen from the flow chart above, in the client side the querying image feature information is extracted and the information index is created. In the server side, the low level visual feature describing data of all the images are stored, and the trained SVM classification model is also stored offline [10]. Under the SVM classification, the querying image and the images in the database are classified based on their vectors. After the classification, the images in the same class with the querying image are the retrieval results [11].

In this paper, this image retrieval module is used.

4. The Realization of the System

4.1. The Construction of the System

In the process of the recognition, the mobile phone is responsible for the image upload and the result display. The server side is responsible for the feature extraction, image recognition, geometric consistency check, and other large computational task. The performance requirement of the phone is low. The construction of this system is shown in figure 5.

In order to recognize large amount of objects, the scalable recognition algorithm based on VT (vocabulary tree) is used [12]. In order to enhance the recognition precision, the SIFT (scale invariant feature transform) feature extraction algorithm is employed [13]. The SIFT algorithm against rotation, scaling, translation, is robust against the change of light, affine transformation, 3D projection in a certain extent.

We can see from figure 5, firstly, visitor uses mobile terminal to collect image and then upload it. If the visitor is interesting in an exhibiting painting, he can take photo and upload it. Then the server receives the image and extracts the SIFT feature. The VT image recognition engine is used and then complete the geometric consistency check. The retrieval work is carried out and the corresponding information is sent back to the visitor. In the visitor’s mobile phone, the information is displayed.

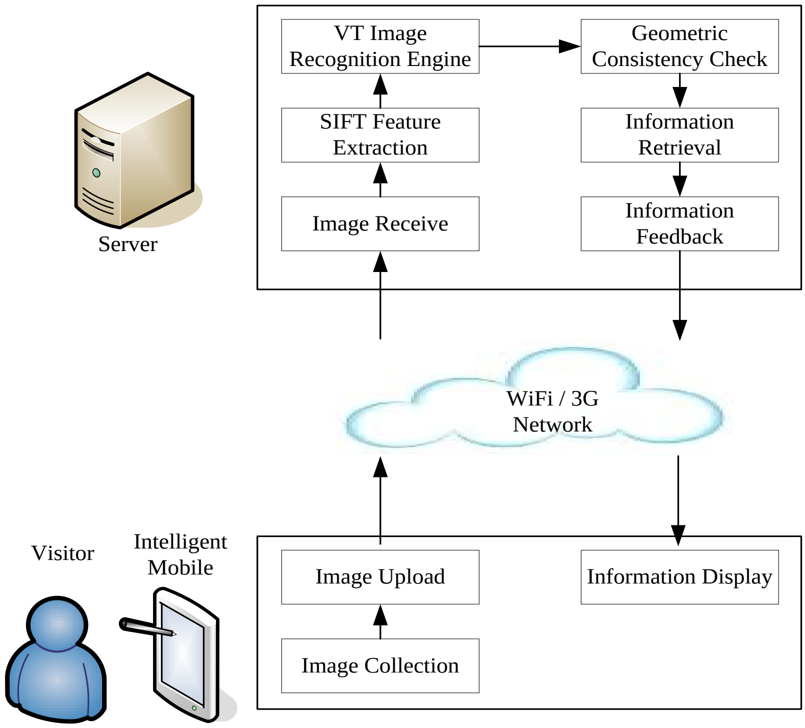


Figure 5. The Guiding System Construction

4.2. Scale Invariant Feature Transform

SIFT (Scale Invariant Feature Transform) is a kind of computer vision algorithm, which is used to detect and describe the local features of the image. It is used to find the extreme points in space scale and extract its position, scale and rotation invariant.

This algorithm is published by Lowe David in 1999, and is summarized in 2004. Its application range includes object identification, robot map perception and navigation, image stitching, 3D model establishment, gesture recognition, image tracking and motion comparison [14].

The description and detection of local image features can help identify objects, and the SIFT features are based on some local appearance of the object. The SIFT features have nothing to do with the size and rotation of the image and can tolerate the light, the noise, some change of visual angle in a certain degree. Based on these characteristics, they are highly significant and relatively easy to be captured. It is easy to be identified with little mistake in the huge amount of the feature database. The detection rate to the partially hidden objects using SIFT features is also very high, and even only more than 3 SIFT object characteristics are sufficient to calculate the object position and orientation. Under the condition of current computer hardware speed and small feature database, the identification speed can be close to the real time operation. SIFT features have huge amount of information and are suitable for rapid and accurate matching in massive database.

4.3. Fast Quantization of the Image Feature

When an image is retrieved using SIFT features, the most simple method is that each image is represented by one feature set. When retrieving, the querying image is compared with each image feature set in the database. The image with the most matching features is returned back. This method is suitable for the database with fewer amounts of images. When the images in database are added with exponential speed, the computing time also increases with exponential speed. So this method is not suitable for the practical application. Nister draw on the experience of the idea of text information retrieval, which hierarchically clusters the image features in the training set [15]. A vocabulary tree is built to quantize the features fast. The steps are described below. Firstly, the SIFT features are extracted from each painting in the training set, gaining the features set *F*. Then the *F* set is clustered hierarchically using *k*-means algorithm. In the beginning, one *k*-means algorithm is used to classify the set *F* into *k* parts, and *k* clustering sets F_i is produced. Then each set F_i is clustered into *k* parts once again, *k* clustering parts $F_i^{'}$ are produced. After *L* times' iteration, a vocabulary tree with *L* levels and *k* subsets comes into being. The leaf node is visual word, and the number of leaf node is k^L. So there is no need to traverse all the visual words when quantizing the feature, and the quantization time is shorten dramatically.

4.4. Huge Amount of Images Retrieval

The weight of the visual word node *i* in vocabulary tree is defined as

$$w_i = ln\frac{N}{N_i} \tag{1}$$

In the formula, *N* is the total number of the database images. N_i is the image number with visual word *i* in the database.

The weight of visual word i in querying image is q_i, the weight of visual word *i* in database is d_i.

$$q_i = n_i w_i \tag{2}$$

$$d_i = m_i w_i \tag{3}$$

n_i is the times that visual word *i* appears in querying image *q*, m_i is the times that visual word *i* appears in database image *d*. If one visual word does not appear in the image, then the weight is set to 0.

The querying image *q* and database image *d* can be represented as the below vectors.

$$q = [n_1 w_1, n_2 w_2, ..., n_\tau w_\tau] \tag{4}$$

$$d = [m_1 w_1, m_2 w_2, ..., m_\tau w_\tau] \tag{5}$$

In the formula above, τ is the number of visual words.

The similarity between querying image and database image is *S(q,d).*

$$S(q,d) = \left\| \frac{q}{\|q\|} - \frac{d}{\|d\|} \right\| \tag{6}$$

If the similarity is calculated via L_2-norm, then the formula above can be simplified.

$$\| q - d \|_2^2 = 2 - 2 \sum_{i|q_i \neq 0, d_i \neq 0} q_i d_i \tag{7}$$

Each time the top *n* similar images are selected as the results, *n* is set to 3 in this paper.

4.5. Geometric Consistency Check

When the *n* most matching results return to users, maybe there is no one is correct. So the users get the wrong identification results. In this paper, geometric consistency check is used to choose the best matching image.

At first, the features are matched between the querying image and candidate images. Then the RANSAC (random sample consensus) algorithm is used to remove the miss matching point [16]. The painting with most matching points is taken as the recognition result. If the matching points are lower than threshold, then there is no matching item in the painting database. In this paper, the threshold is assigned to 3.

5. The Mobile User Interface

The so-called mobile UI (User Interface) refers to connect the concept and visual elements through the exhibits in the Picture Gallery. Through the effective design of mobile application, the thought of the visitors and the art works are combined to strengthen the visitors' sense of identity. The unified color tone, the unified structure, and the unified interactive behavior are helpful to shape visitors' aesthetic idea. The exquisite dynamic methods in the mobile application can improve the display effects of the exhibiting arts. The abstract concept demand is transferred to visual effect after the visitors receive the response from the server. This method optimizes the users' experience greatly and shortens the distance between the visitors and the exhibits, so the visitors will have a sense of identity [17].

The system based on the user's mobile terminal and the Picture Gallery server can improve the users' experience on the exhibiting paintings. The exhibits can be displayed on the visitors' mobile phone UI interface vividly. The mobile terminal can connect to the Picture Gallery server via 3G or WiFi network. The server side can determine the corresponding exhibits through the visitors' inputting information and the affluent information about the paintings can be shown on the mobile UI via wireless network. The mobile can communicate with the server without any other auxiliary devices. The server can optimize the user interface and call out the information which satisfies the visitors' requirement from the database. The corresponding information is shown on the visitors' mobile and caters to their requirement.

The main interface of the system is shown in figure 6. The title of this mobile application is “Mobile Visual Search Guiding System”, and the background of the mobile application is the Da Vinci's painting-Mona Lisa, under which is a “camera image” button. When pressing the the camera button of the application, the photo-taking function of the mobile phone is enabled and the interesting painting on exhibition will be photographed and sent to the server in Picture Gallery to be retrieved.

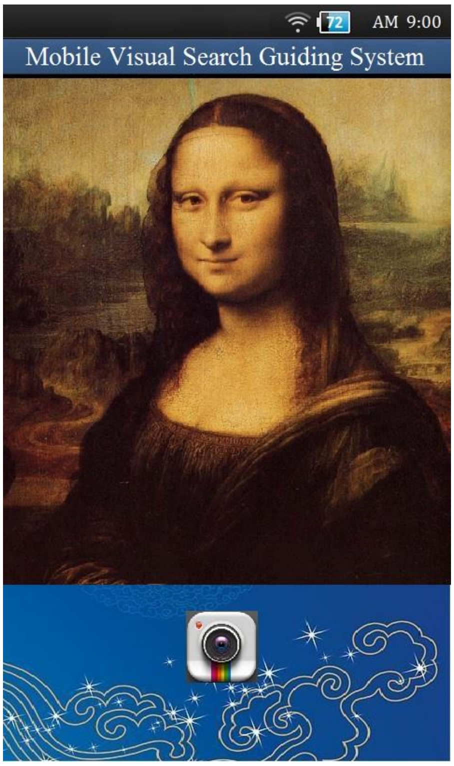


Figure 6. The Main Interface of the System

6. Experimental Results and Analysis

In this paper, a high performance PC is selected as the server, which is equipped with 3.9 GHz CPU, 8 G RAM. The mobile phone is SAMSUNG GALAXY S5 G9008V, equipped with 2.5GHz CPU, 2 GB RAM. The wireless network is 54 Mbps WiFi.

In order to estimate the performance of the recognition, the Standard Object Recognition Data Set is chosen [18]. In the data set, there are books, CD boxes, toys and so on. Each object has 4 pictures in different perspectives and illumination. The main aim of the test is to find out how far does the number of the identifying objects affects the recognition precision and retrieval speed.

In this paper, 250 objects are used to test. Each object has 4 pieces of pictures, so there are 1000 pieces of pictures in total. Every 4 pieces of pictures, 3 of them is for training and 1 of them is for testing.

The degree that the number of the objects affects the recognition precision is shown in figure 7. We can find that the bigger of the objects number, the lower of the recognition ratio. When there were 50 objects, the recognition precision was about 91%. When the objects were increased to 250, the recognition precision was about 69%. So the geometric consistency check is used to reduce the false recognition rate.

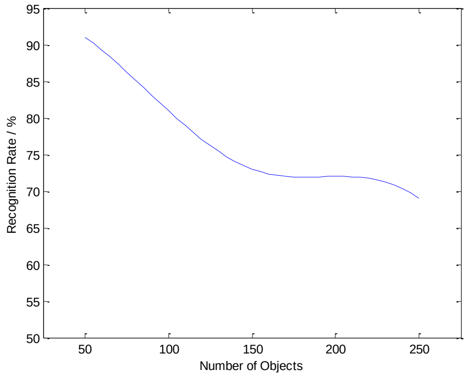


Figure 7. Effect That Number of the Objects to Recognition Precision

The degree that the number of the objects affects the recognition speed is shown in figure 8. It is obvious that the consuming time increased little when the number of the objects was added up to 250. When the number of the objects changed from 50 to 250, the consuming time only increased about 0.02 seconds. So the VT algorithm adapted in this paper is suitable for huge amount of objects recognition.

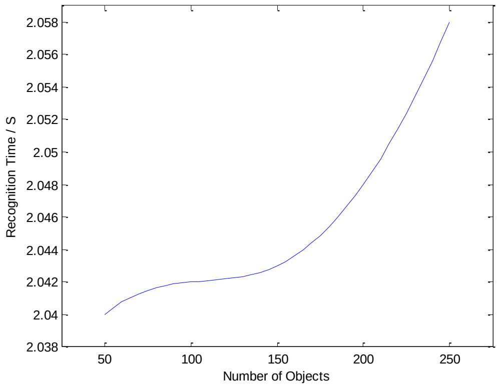


Figure 8. Effect That Number of the Objects to Recognition Time

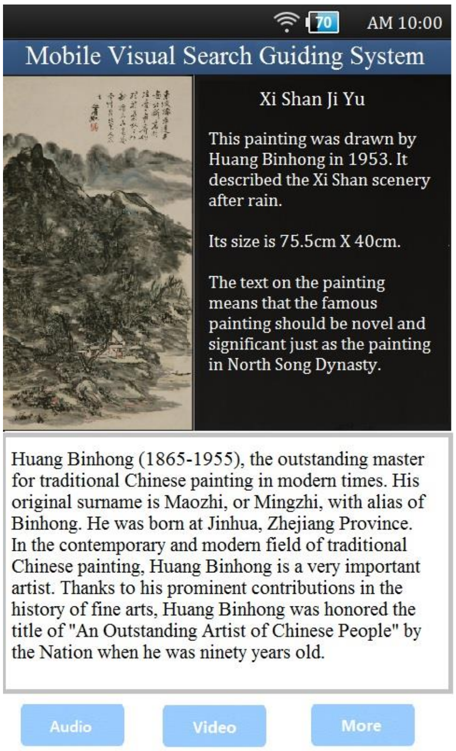
The processing time of 250 objects recognition is listed in table 1. It is obvious that the SIFT feature extraction costs the most of the time, about 97.8%.

Table 1. Processing Time of the Recognition Algorithm

	Average Processing Time/ms	Occupancy Rate/%
SIFT Feature Extraction	2521	97.8
VT Recognition	56	2.2

As the experiment shows, the precision and speed of the recognition algorithm adopted in this paper is suitable for Gallery guiding system. When an exhibiting paiting is recognized, it costs about 3 seconds.

Limited by time and condition, we took a small scare test in National Art Museum of China. If more paintings take into account, there should be more training work. When we took a photo of the famous painting *Xishan Jiyu* with a SUMSUNG mobile phone, the result was shown in phone as figure 9. We can gain the information about the art and its author. The interface is divided into 3 parts. The painting is shown in the top left. The top right is the introduction of this painting. The information about painting artist is shown in the lower part. There are 3 buttons in the bottom of the interface, "Audio", "Video", "More". If more information is needed, the audio and the video about this painting can also be given. As the experiment result shows, the mobile visual search guiding system is feasible and practical.



Mobile Visual Search Guiding System

Xi Shan Ji Yu

This painting was drawn by Huang Binhong in 1953. It described the Xi Shan scenery after rain.

Its size is 75.5cm X 40cm.

The text on the painting means that the famous painting should be novel and significant just as the painting in North Song Dynasty.

Huang Binhong (1865-1955), the outstanding master for traditional Chinese painting in modern times. His original surname is Maozhi, or Mingzhi, with alias of Binhong. He was born at Jinhua, Zhejiang Province. In the contemporary and modern field of traditional Chinese painting, Huang Binhong is a very important artist. Thanks to his prominent contributions in the history of fine arts, Huang Binhong was honored the title of "An Outstanding Artist of Chinese People" by the Nation when he was ninety years old.

Figure 9. The Recognition Result

7. Conclusions

In this paper, the SVM was used for the classification training and the SIFT feature extraction algorithm was used to enhance the recognition rate. Then the geometric consistency check was employed to choose the best matching image. Finally the mobile visual search guiding system was implemented.

The guiding system proposed in this paper can be widely used in the Picture Gallery. Visitors can use mobile phone to take a photo of the painting on show to obtain the detail information. Compared with other guiding system, the system proposed in this paper has the following advantages.

The computing visual method is used to identify the showing paintings, which is a humanized design and there is no need to input identifier or keyword. There is no need to install auxiliary devices in the Picture Gallery and it is suitable for wide popularization.

The detail information of the exhibiting paintings stores in the server, so there is little requirement for the client side's storage capacity. The data updates in the server side and the operation is more convenient. When the server completes the information update, all the clients can get the latest information.

For the Picture Gallery, there is no need to hire any other devices. The only work to do for the visitors is to install this application on their intelligent phone. The application can afford text, voice, video and other type of information to the visitors.

As the experimental results show, it is feasible that the mobile visual search technology is applied to the Picture Gallery guiding system. With the development of the network technology and the popularity of the intelligent phone, this method will have a wonderful future.

The future work is to explore the fast, robust image feature extraction algorithm to enhance the retrieval speed and the recognition precision. This system will be more real-time and have better user experience.

Acknowledgements

This work is supported in part by Projects of Shandong Province Higher Educational Science and Technology Programs (J12LN31, J13LN11, J14LN59), the Development Projects of Science and Technology of Shandong Province (2012GGE27073, 2014GGX101011, 2015GGX101018), the Shandong Xiehe College School Fund (XHXY201431).

References

[1] Y. Bizhong, "The Interaction Design and Implementation of the Guided Tours System Based On The Smart Phone", Beijing: Beijing University of Technology, **(2014)**, pp. 3-11.

[2] S. Vosinakis and P. Koutsabasis, "Interaction design studio learning in virtual worlds", Virtual Reality, vol. 17, no. 1, **(2013)**, pp. 59-75.

[3] B. Girod, V. Chandrasekhar and R. Grzeszczuk, "Mobile Visual Search: Architectures, Technologies, and the Emerging MPEG Standard", Multimedia IEEE, vol. 18, no. 3, **(2011)**, pp. 86–94.

[4] E. Bruns, B. Brombach and T. Zeidler, "Enabling mobile phones to support large-scale museum guidance", IEEE Multimedia, vol. 14, no. 2, **(2007)**, pp. 16-25.

[5] Y. Liu, "A survey of content-based image retrieval with high-level semantics", Pattern Recognition, vol. 40, no. 1, **(2007)**, pp. 262-282.

[6] Z. P .Shi, H. He, Y. Liz, Z. Sic and L. Duan, "Texture spectrum descriptor based image retrieval", Journal of Software, vol. 16, no. 6, **(2005)**, pp. 1039−1045.

[7] L. Wencheng, Y. C. Chang and H. H, Chen, "Integrating Textual and Visual Information for Cross-language Image Retrieval", Information Processing and Management, vol. 43, no. 5, **(2007)**, pp. 488-502.

[8] X. Qi and Y. Han, "Incorporating multiple SVMs for automatic image annotation", Pattern Recognition, vol. 40, no. 2, **(2007)**, pp. 728-741.

[9] R. Liu, Y. Wang and T. Baba, "SVM-based active feedback in image retrieval using clustering and unlabeled data", 12th International Conference, CAIP 2007, Vienna, Austria, **(2007)**, pp. 954-961.

[10] J. Cheng and K. Wang, "Active learning for image retrieval with Co-SVM", Pattern Recognition, vol. 40, no. 1, **(2007)**, pp. 330-334.

[11] C. Cusano, G. Ciocca and R. Schettini, "Image annotation using SVM", Proceedings of the SPIE, vol. 5304, **(2003)**, pp. 330-338.

[12] Z. Y. Hui, W. Jian, L. S. Shan and C. Z. Ming, "Algorithms of Vocabulary Tree Hierarchical Semantic Model Image Retrieval", Microelectronics & Computer, vol. 29, no. 11, **(2013)**, pp. 172-176.

[13] G. Xu, Z. Zhang, W. Qi, M. Liao, L. Xu and H. Zhao, "Image Automatic Annotation Based on the Similarity of Regions", Journal of Computational Information Systems, vol. 10, no. 21, **(2014)**, pp. 9397-9404.

[14] Z. H. Huiyu, Y. Yuan and C. Shi, "Object tracking using SIFT features and mean shift", Computer Vision & Image Understanding, vol. 113, no. 3, **(2009)**, pp. 345-352.

[15] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, Washington DC: IEEE Computer Society Press, vol. 2, **(2006)**, pp. 2161-2168.

[16] M. A. Fishler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography", Communications of the ACM, vol. 24, no. 6, **(1981)**, pp. 381-395.

[17] X. X. Liu, J. Y, Zheng, "UI Design of Touch Screen Mobile Phone", Packaging Engineering, vol. 30, no. 2, **(2009)**, pp. 130-132.

[18] B. Girod, V. Chandrasekhar and D. M. Chen, "Mobile visual search: linking the virtual and physical worlds", IEEE Signal Processing, vol. 28, no. 4, **(2011)**, pp. 61-67.

Authors

Gongwen Xu, he works in School of Computer Science and Technology Shandong Jianzhu University, and received the Master's degree in 2005 from Shandong University. His interesting fields are spam detection, image processing, information retrieval.

Xiaomei Li, she is a member of Cancer Center of the Second Hospital, Shandong University. Her research interests include medical image processing and pattern recognition. She received the Master's degree in 2004, M.D. degree in 2014 from Shandong University.

Jian Lei, he works in Shandong Institute of Standardization, Dongying Branch. His research interest is image processing.

Honglan Zhou, she works in Dongying Special Equipment Inspection Institute. Her research interest is information retrieval.

Zhijun Zhang, he is working as an assistant professor in School of Computer Science and Technology Shandong Jianzhu University. His interesting filed is pattern recognition.