MVSS: Mobile Visual Search Based on Saliency

Yegang Du Zhiyang Li Wenyu Qu Siyang Miao Songhe Wang
College of Information Science and Technology
Dalian Maritime University
Dalian 116023, P.R.China
Email: duyegang@gmail.com

Heng Qi
School of Computer Science and Technology
Dalian University of Technology
Dalian 116024, P.R.China
Email: qhclement@gmail.com

***Abstract*—With development of content-based image retrieval (CBIR), mobile visual search (MVS) is a promising application. In typical MVS, similar images are retrieved from the database maintained by the server, given a query image taken by mobile devices. Different from general CBIR, the problem of transmission latency should be considered in MVS. In existing work, the progressive transmission is proposed to minimize the data size in transmission by low-dimensional feature descriptors and compression coding in order to reduce the transmission latency in MVS. Although the retrieval speed is improved by existing progressive transmission methods, the result accuracy is decreased because of the information loss in these methods. To address this problem, this paper proposes a novel framework for MVS which consists of a new progressive transmission model based on image saliency (MVSS) and a new distance metric corresponding to the proposed progressive transmission model. In our framework, we use SIFT descriptors to represent images, which can preserve more information than other low-dimensional feature descriptors and compression coding. Although SIFT is high-dimensional descriptor, we only transmit the SIFT descriptors in salient regions of image to reduce the transmission latency. We evaluate our framework on Stanford image set, and the results demonstrate that our framework not only reduces the transmission latency but also achieves a better retrieval accuracy.**

***Index Terms*—saliency; bag of words; distance algorithm**

I. Introduction

In recent years, mobile terminals including smart phone and tablet PC have been developed greatly in areas of communications and multimedia [1] [2]. These devices are equipped with high-resolution camera, touch screen and high-performance CPU, which provide a perfect platform for Mobile Visual Search (MVS). So MVS applications typically follow a client-server architecture. The mobile client takes a query photo which is then transmitted to the server. According to the content received, the server retrieves relevant information from its maintained photo database, and gives the feedback to the client mobile at last. In this process, some technology could be used in the server-side like [3] [4]. Although some applications have come into services, there are still some challenges on mobile visual search. How to decrease the responding latency becomes one of the biggest challenges in both industry and academia [5].

Generally, the responding latency can be divided into three parts: client-side processing, transmission and server-side processing [6]. Due to the data is transmitted via a bandwidth constrained wireless link, the transmission becomes the most time consuming part as depicted in Fig. 1. In that case, The performance of a MVS system heavily depends on how much data transmitted over the network. Many methods concentrated on how to reduce the size of transmitted data. Sending feature vectors or descriptors instead of captured images is proved to be efficient [5]. Thus, typical MVS system flow could be summarized as shown in Fig. 2. Various presented image descriptors can be used in the feature extraction process.

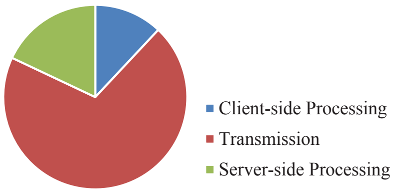


Fig. 1. Component of latency

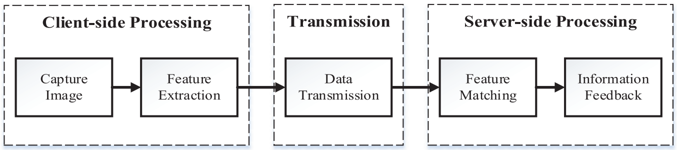


Fig. 2. MVS flow

The state-of-art visual descriptor is scale invariant feature transform (SIFT). However, the size of SIFT descriptor is larger than other ones, even larger than the captured image sometimes, which make it inappropriate to be directly used in MVS system. Thus, many subsequent researches focus on how to reduce the 128-dimensional SIFT descriptor into lower dimensional space, such as Locality Sensitive Hashing (LSH) SIFT [7], Similarity Sensitive Coding (SSC) SIFT [8] and Principal Component Analysis (PCA) SIFT [9]. There are also some other low-bit rate descriptors such as: Speeded-Up Robust Features (SURF) [10], Gradient Location and Orientation Histogram (GLOH) [11] and Compressed Histogram of Gradients (CHoG) [12].

Another direction to decrease the responding latency is to improve the conventional pipeline by using progressive transmission strategy. The method proposed by Chandrasekhar [13] extracts CHoG features firstly, and then sorts the descriptors by Hessian responses. Those descriptors with higher

Hessian response have higher priority. The server begins to match the features when the first several descriptors arrive. However, this method loses sight of the geometric information in the captured image. In [6], the authors propose a geometric context-preserving progressive transmission method by dividing the captured image into several blocks and features in same block are transmitted as a unit. Compared to the previous transmitting strategy [13], the proposed strategy increases the retrieval rate, meanwhile reduces the responding latency to some extent. However, coherent regions might be segmented into different blocks sometimes, which is contrary to the mechanism of human visual systems.

Given an image, human visual system tends to find the most informative regions and analyze the content of these regions, which are usually called salient regions in computer vision. Thus, this paper proposes a progressive transmission strategy based on image saliency. Upon capturing the image, we compute the saliency value of each pixel by the spectral residual approach [14] and divide the pixels into different saliency levels. After that, the feature points on the image are detected and the corresponding feature vectors are computed and then transmitted to the server. We choose SIFT descriptors in the experiment since SIFT is the most matured one and its implementation is free on the Internet. To decrease the responding latency, the computing and transmitting process proceed in order of the saliency levels of the feature points, and will break if an appropriate result is retrieved.

Our main contributions can be summarized in three parts:

- Compared to the previous progressive transmission methods [13], [6], our presented strategy follows the mechanism of human visual system, which can preserve the coherent regions in the image as a whole and effectively reduce the responding latency.
- To compare the silent regions with images in the database, the extracted local features are encoded into global histograms via Bag of Words (BoW) framework. For improving the matching rate, we present a new distance measure which considers different magnitudes between the histograms encoded from these two kinds of images.
- Extensive experiments on the Stanford image data set have shown that the proposed MVSS system generally outperforms the baseline method (sending all the feature vectors), since retrieving by salient regions can alleviate the influence of the background of the image during the retrieval process.

II. Descriptor transmission model based on saliency

As expounded in [6], points in the same object usually have the similar feature information. And block-by-block transmission is an effective method to reduce the size of data transmission and keep the geometric context information [5]. But dividing the captured image using grid lines roughly may lead to some mistakes shown in Fig. 3. Points in the same object are divided into three blocks, which will break the geometric context. According to the features of human visual system, we propose a more reasonable transmission model using the saliency of the captured image.

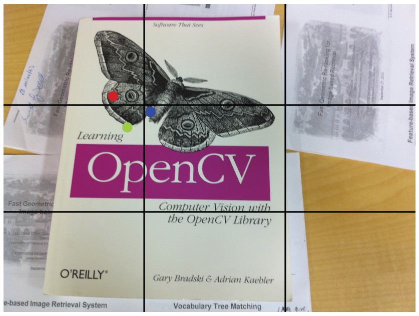
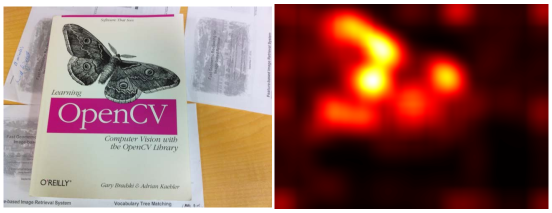


Fig. 3. Mistakes caused by division

Actually, various kinds of saliency detection methods are presented in the literature [15], which are related to many applications, such as object recognition, automatic image cropping, image/video compressing. In this paper, we choose the saliency detection based on spectral residual (SR) presented in [14], due to its fastness and easiness to implement on a mobile terminal. The saliency map of the original image shown in Fig. 4 can be obtained by transforming spectral residual into spatial domain. The saliency map is used to determine the different levels of importance for the content in the image. We will give a detailed description of the proposed MVSS system in the following.



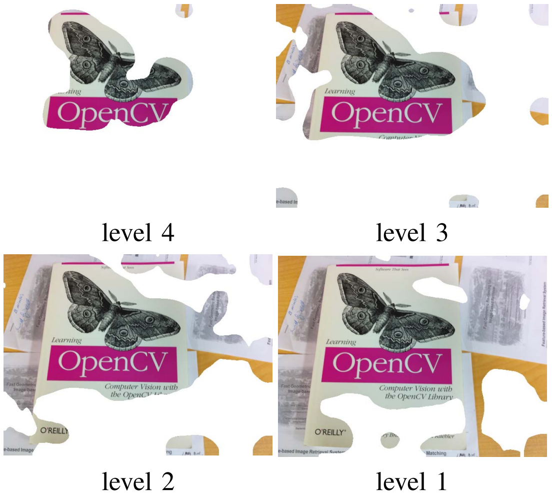
(a) (b)

Fig. 4. The original image (a) and its saliency map (b).

A. Client-side

Once obtaining the captured image, the saliency values are computed by the mentioned SR approach for each pixel on the image. That is to say, we have a saliency matrix with the same size as the image. Then M levels of salient regions like a terrace are determined by selecting different quantiles as shown in Fig. 5. For example, we set level 1 represents the quantile of 20% and level 2 represents the quantile of 40%, and the rest can be done in the same manner. Then from the figure below, we can see that different levels generally mean different significance. Contents in several most salient regions are the most sensory information of the image.

Upon obtaining the saliency values of the query image as described above, the next step is computing SIFT descriptors. SIFT computation consists of two steps: feature points detection and feature vector computation. As the first step is fast


level 4 level 3

level 2 level 1

Fig. 5. Different levels of salient regions.

and the second one is somehow time consuming, we detect all the feature points on the image at once and compute the feature vector for each point in order of its saliency level. That is to say, the features for points with higher saliency values are computed before the ones with lower values.

The server starts or restarts the retrieval process when it receives a complete group of features. Generally speaking, there is no need to transmit all the features before we get the right feature matching. So our transmission mechanism can greatly reduce the amount of transmitted data, and then decrease the transmission latency.

B. Server-side

Due to the high performance of the server-side, we choose the BoW framework [16] to do the feature matching. The BoW framework utilizes the inverted index method which is widely used to retrieve large image databases rapidly and effectively. And this framework can be described in Fig. 6.

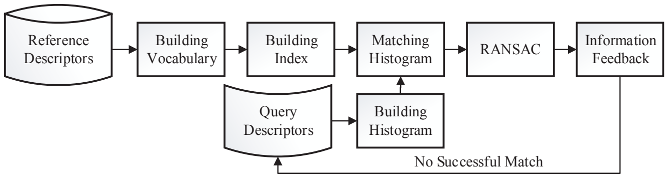


Fig. 6. The bag of words framework

First, SIFT descriptors are extracted for all the reference images in the database and later put into a 128-dimensional space. Then, the vocabulary or codebook is constructed by clustering the descriptors into k words. Here we use Approximate K-means (AKM) in the experiment since it can handle large data set. The size of codebook k is generally set based on the scale of the descriptors to get the best result. As it is pointed that, codebooks with smaller size always have less distinguish ability and the ones with larger size will contain noise words.

The second step is building the index. Inspired by text retrieval, the reference images can also be inverted by its visual words. Thus, when retrieving images, a query image q only needs to match with the reference images inverted by the words occuring in q, which will heavily reduce the matching time. To efficiently match the query image and a reference image, both images are represented as vectors of word frequencies in the codebook. This is generally called encoding process which encodes the local features into a spatial global histogram.

To improve the accuracy of the distance between two histograms, normalization of the histograms should be done firstly. Term frequency-inverse document frequency (tf-idf) normalization is a state-of-art method. Specifically speaking, if a vocabulary contains K words, then the weighted word frequency is computed as followed:

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \tag{1}$$

A detailed discuss of these parameters are included in [16]. The normalization includes all the reference images in the database. Although the first two steps are time consuming, they only need to be carried out offline, which means the process would not produce any latency during retrieval.

Once a group of features F are received from the mobile terminal, F will be encoded into a global histogram q according to the pre-built codebook. Then q is normalized by Eq. 1. After that, the similarity of the query image and the reference image can be represented by the distance between their corresponding histograms. The traditional distance measure includes L_1 distance, L_2 distance, *cos* distance and so on. Since F is only a small part of original feature set, the query histogram generally will have less magnitude than the reference histogram. Using the above distance measure directly leads to low accuracy in practices. We present a new distance measure to overcome the above problem.

$$D(q,r) = \frac{1}{\sum_{i \in I} \min(\hat{q}_i, r_i) * \sum_{i \in I} \max(\hat{q}_i, r_i)}, \tag{2}$$

$$\hat{q}_i = \frac{q_i}{\sum_{i \in I} r_i}.$$

where q and r are the query histogram and a reference histogram in the database. Set I is the index set of visual words occuring in q, which obviously is a subset of the index set of the codebook. From Eq. 2, we can see the $D(q,r)$ will reach a local minimum when $\sum_{i \in I} \min(\hat{q}_i, r_i)$ approximates to $\sum_{i \in I} \max(\hat{q}_i, r_i)$, which implies the query histogram q finding a similar part $\{r_i | i \in I\}$ in the reference histogram r. Meanwhile, $D(q,r)$ tends to have a large value as $\sum_{i \in I} r_i$ increases which reflects that r has more features similar to q. The following experiment results in Section III demonstrate the new distance measure outperforms other existing measure in most of cases, and is more suitable to match histogram with different magnitudes.

Finally, the reference images are ranked by distances between them and the query image via Eq. 2. Generally, in order to improve the retrieve rate, some candidate images are chosen and proceeded with spatial checking process. RANSAC is a state-of-art checking method which is also used in this paper to remove the erroneously matched points. After RANSAC checking, we get most similar candidate images to the query image, and the relevant feedback information is sent to users.

III. Experiment and Performance Analysis

Due to the unstable network bandwidth in realistic environment, we choose to do the MVS experiments on PC. The experiments are carried out on a computer with a Intel(R) Core(TM) i3-2130 3.4GHz CPU and 2GB memory. The code is implemented in MATLAB 2012b with some parts written in C with MEX interface. The implementation of SIFT, SR and BOW are based on the VLFeat toolbox [17], Spectral/Phase-based Visual Saliency toolbox [18] and Caltech Large Scale Image Search toolbox [19] respectively, which are all publicly available in the Internet.

Scale and constitution: Stanford data set which is well known in MVS contains over 3269 images and these images cover eight different categories. They are book covers, business cards, cd covers, dvd covers, landmarks, museum paintings, print and video frames. These categories all consist of two parts: query images and reference images. The reference images can be regarded as standard, and the query images to be retrieved are taken by different mobile terminals. There are in total 2677 images captured from mobile phone cameras and 592 images from Canon digital cameras.

Since the transmission latency is difficult to simulate in PC, we can estimate the transmission latency by measuring the amount of data needed to be transmitted. For each category, we regard the query images as images captured by mobile terminal, and the reference images are constituted into a photo database as the sever side maintains. The retrieval rates in each category are computed respectively. The original SIFT+BoW method is chosen to be a baseline method.

In the following, we describe the parameters used in the experiments. For all experiments, each image is converted to gray-scale format with a fixed size (width 640 pixels), keeping the original image aspect ratio. When encoding the local feature vectors to the vocabulary by AKM method, the vocabulary size k is an important parameter. It is pointed that larger vocabularies generally lead to higher accuracy. In our experiment, we set $k = 30K$ as the default case. Upon obtaining the saliency value of each pixel in the image, all the pixels are classified into $M = 5$ levels. In RANSAC spatial checking process, we select $t = 20$ proximal reference images for each captured image to do the RANSAC checking and select the one with the highest score.

A. The performance analysis of the vocabulary size k

The vocabulary size k will directly affect the effectiveness of subsequent indexing process, matching process, and then the final retrieval rate. As is known that a proper k is related to the size of data set. Larger vocabularies tend to have higher accuracy. However, with the growing size of vocabularies, the time complexity of matching process will grow too.

To select a proper k for Stanford data set, we set $k = \{10K, 15K, \ldots, 55K, 60K\}$ and do the retrieval experiments using the baseline method with all the other parameters chosen in the default case. All the results of the eight categories are listed in Table I and the average retrieval rate is showed in Fig. 7. From the result, we can see even when a vocabulary size k is $60K$, the recognition rate still appears to be increasing which suggests that further gains could be achieved by increasing k.

Meanwhile, it should be noted that the average retrieval rate grows fast with k growing before $k = 30K$, and the growth is slower after $k = 30K$. Thus, $k = 30K$ is proper chosen considering both the efficiency and accuracy. So we set $k = 30K$ as the default value in the experiment.

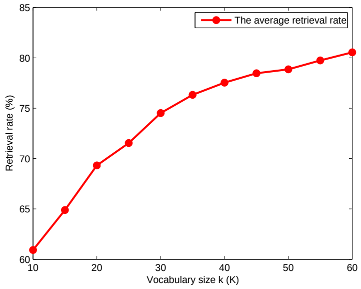


Fig. 7. The chart of average result

B. Analysis on saliency level and transmitting data size

In this experiment, we extract SIFT descriptors from all query images and transmit them in order of their saliency levels. Since the transmission latency heavily depends on the amount of SIFT descriptors transmitted over the Net. Thus, it is interesting to know the relation between a saliency level M_i and the corresponding transmitting data size N_i in this level.

After computing the saliency values of current query image, we equally divide the saliency values into 10 different levels indicated as $M_i, i = 0, \ldots, 9$. Thus, the image pixels are segmented into 10 non-overlapping regions based on the pixels' saliency levels. Let N and N_i represent the number of SIFT feature points in the query image and i-th segmented regions respectively. Noticing the region may not be a connected area, but it is easy to count the number of SIFT feature points in each region. N_i can be used to measure the amount of transmission data in the saliency level M_i.

For each query image in Stanford data set, the transmission ratio of descriptors in each saliency level M_i is computed by Eq. 3

$$R_i = \frac{N_i}{N}, \tag{3}$$

The average transmission ratio in each saliency level among all the images are illustrated in Fig. 8, which shows that the relationship of saliency level and corresponding R_i are

TABLE I
VOCABULARY SIZE AND RETRIEVAL RATE

	10K	15K	20K	25K	30K	35K	40K	45K	50K	55K	60K
book covers	86.39%	89.85%	93.07%	94.06%	95.30%	95.79%	96.29%	96.29%	97.03%	96.78%	96.78%
business cards	51.25%	53.75%	62.50%	66.75%	72.50%	76.00%	76.50%	78.25%	79.50%	82.00%	83.25%
cd covers	67.75%	72.75%	79.75%	80.00%	86.75%	87.75%	90.25%	91.00%	91.25%	93.50%	93.75%
dvd covers	75.75%	81.25%	84.50%	86.25%	88.00%	88.25%	88.00%	88.50%	89.25%	88.75%	88.75%
museum paintings	68.96%	72.80%	76.37%	76.10%	80.22%	81.04%	82.97%	84.07%	84.34%	86.26%	86.81%
print	38.75%	42.00%	45.00%	50.75%	55.00%	60.50%	63.25%	63.25%	63.25%	63.25%	63.25%
video frames	73.75%	78.75%	86.25%	89.25%	89.25%	92.00%	93.50%	95.25%	95.50%	96.75%	96.50%
landmarks	24.75%	27.94%	27.15%	29.14%	29.14%	29.34%	29.54%	31.14%	30.74%	30.74%	35.33%
Average	60.92%	64.89%	69.32%	71.54%	74.52%	76.33%	77.54%	78.47%	78.86%	79.75%	80.55%

approximately linear. That is, when the saliency level turns higher, the corresponding R_i will low down with similar portion.

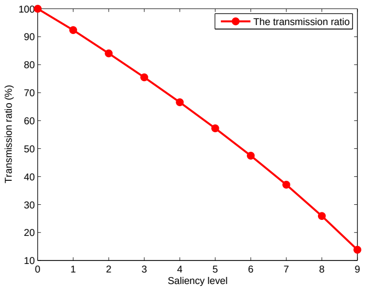


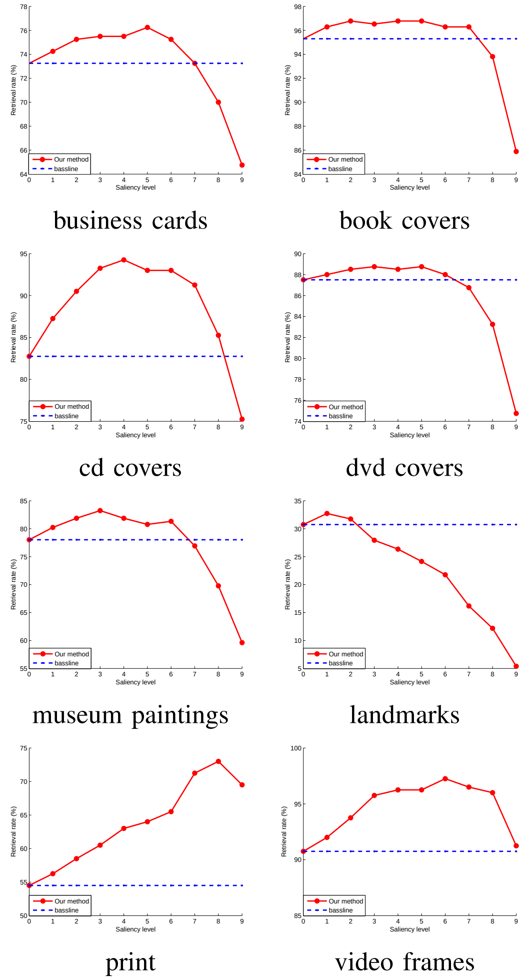
Fig. 8. The mean value of Saliency level and corresponding R_i

This experiment demonstrates that saliency level is inversely proportional to the quantity of transmitted descriptors, and our method can indeed reduce the data size transmitted on the network. For instance, if the server find the right match with level 9, our method can achieve nearly 90 percent reduction on transmission latency.

C. Retrieval rate and saliency level

In this section, we compare our method with the baseline method (SIFT + BOW) in Fig. 9 and Fig. 11. Although saliency level is chosen as 5 in MVSS system, we set saliency level to 10 in the experiments to see the relationship of retrieval rate and saliency level. The vocabulary size k is set to $30K$, and all the other parameters are chosen in default cases. The abscissas represent saliency level M which means transmitting the SIFT descriptors in the first M saliency levels to do the retrieval experiment. The ordinates represent the corresponding retrieval rate.

From Fig. 9, we can see that our method performs better than the baseline method on seven categories of data set. However, it is less effective on landmarks. With more experiments, we notice that our method does not perform well on landmarks not only because of the bad performance of the baseline method but also because the saliency regions are not extracted correctly. This phenomenon is called ”saliency reversal” problem in [20]. Fig. 10 is an instance of ”saliency reversal”. Cars are the main body of the selected regions in the first three silence levels. However, the actual retrieval object is the building. We will optimize our method to overcome this problem in the further research.



business cards

book covers

cd covers

dvd covers

museum paintings

landmarks

print

video frames

Fig. 9. The contrast between our method and baseline method

Fig.11 shows the average result on the Stanford data set. From which, we can see that the retrieval rate of our method firstly increases and then decreases, with the increase of the silence level M. The reason is that when M is higher than 6, with the increase of M, the salient regions become smaller and the descriptors transmitted to the server become less too.

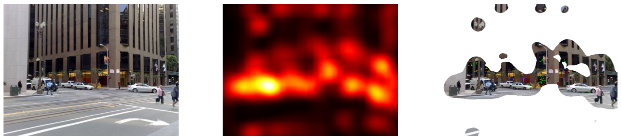


Fig. 10. The "saliency reversal" problem. Take saliency level 7 as examples. This picture is chosen from landmarks on Stanford

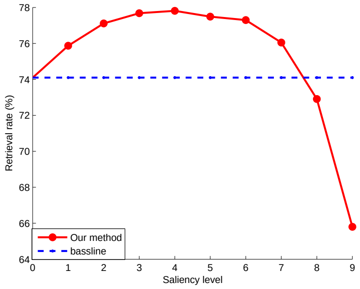


Fig. 11. The average result

Then less descriptors bring lower retrieval rate. However, if right matches are not achieved at level 6, with silence level M decrease further, the noisy descriptors in the background are transmitted. Then the retrieval rate will decrease. We can see that when M is set to 6, the retrieval rate is much higher than the baseline method. Thus, after transmitting the descriptors in first four silence levels, we will probably obtain the right match. In fact, the descriptors transmitted for an image are only one third or less of its entire descriptors.

D. The proposed distance measure

To further discuss the performance of our distance measure defined in Eq. 2, we compare this new measure with the most commonly used measures: L_1 distance, L_2 distance and cos distance. When matching the histograms, many distance measures are presented in the literature. To make a fair experiment, all the other parameters are chosen in the default case. The only difference is the normalization method before the computation of each distance. When computing L_2 distance and cos distance of a query histogram and a reference histogram, both of the histograms are preformed with L_2 normalization first. Meanwhile, L_1 normalization is chosen for L_1 distance and our distance. The average retrieval rates on the eight categories utilizing the different distance algorithms are shown in Fig.12.

From the bar graph, we can see that our distance measure make the retrieval rates higher as silence level M increase when M is higher than 2, on the contrary, make the retrieval rate lower as M decrease compared with other three distances when M is lower than 2. This means that when only a smaller part of the entire descriptors is encoded, the proposed distance tends to achieve higher retrieval rate and preform better than other algorithms. Thus, the proposed distance is more likely to find the right match with a small number of feature descriptors

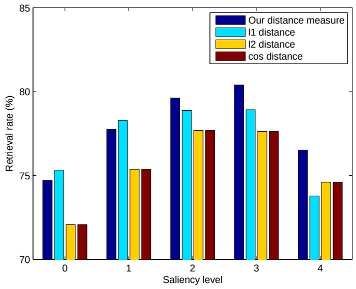


Fig. 12. The average result

transmitted at the first several silence levels. In this case, our distance algorithm is proved to be more appropriate for our saliency transmission method.

E. The comprehensive experiment

Finally, we combine our saliency transmission method with the new distance algorithm as the final MVSS system to do some experiments, and to show the overall performance of the MVSS system. We compare our MVSS system with the original SIFT + BoW method (baseline), Salience SIFT + BOW method. L_1 distance is used in these latter two methods due to its outperformance over other distance measures in experiments. And our distance measure as introduced above is used in our MVSS system. RANSAC is also utilized for all the three methods here to improve the retrieval rate. The data set of landmarks is proved to be inappropriate to our MVSS system in the experiments before, so we do not take it into consideration in this part. The results are shown in Table.II III IV and Fig.13.

TABLE II
MVSS SYSTEM

Saliency level	0	1	2	3	4
book covers	95.30%	96.29%	97.28%	97.03%	94.55%
business cards	94.50%	95.50%	96.00%	94.75%	87.75%
cd covers	93.00%	94.75%	95.50%	95.25%	91.50%
dvd covers	88.00%	89.00%	89.25%	89.00%	85.75%
museum paintings	92.58%	94.23%	94.51 %	95.88 %	93.41%
print	57.25%	59.00%	67.75%	69.50%	73.00%
video frames	93.75%	96.50%	96.25%	97.25%	97.00%
Average	87.77%	89.32%	90.93%	91.24%	88.99%

TABLE III
ORIGINAL SIFT + BOW

Saliency level	0	1	2	3	4
book covers	94.55%	94.55%	94.55%	94.55%	94.55%
business cards	94.00%	94.00%	94.00%	94.00%	94.00%
cd covers	92.50%	92.50%	92.50%	92.50%	92.50%
dvd covers	88.00%	88.00%	88.00%	88.00%	88.00%
museum paintings	93.41%	93.41%	93.41%	93.41%	93.41%
print	65.00%	65.00%	65.00%	65.00%	65.00%
video frames	93.00%	93.00%	93.00%	93.00%	93.00%
Average	88.64%	88.64%	88.64%	88.64%	88.64%

TABLE IV
SALIENCY SIFT + BOW

Saliency level	0	1	2	3	4
book covers	94.55%	95.79%	96.78%	95.79%	94.06%
business cards	94.00%	94.25%	94.75%	94.75%	86.50%
cd covers	92.50%	94.50%	95.00%	94.25%	90.75%
dvd covers	88.00%	89.25%	89.25%	88.75%	85.75%
museum paintings	93.41%	93.68%	94.51%	95.05%	94.23%
print	65.00%	65.50%	70.00%	72.75%	72.00%
video frames	93.00%	94.75%	95.25%	96.25%	97.25%
Average	88.64%	89.67%	90.79%	91.08%	88.65%

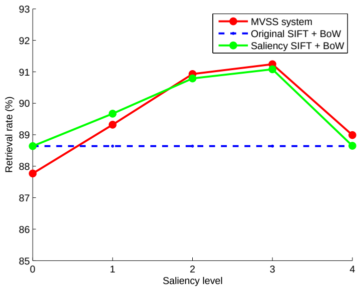


Fig. 13. The average result

As depicted in Table.II, the retrieval rate of our method gets further improved with RANSAC. The retrieval rates on those categories are all higher than the original SIFT + BoW method and the highest retrieval rate can reach as high as 97.28%. And the average result shows that when the saliency level is level 4, the retrieval rate is a little higher than the original SIFT + BoW method. When the saliency level is level 3, our MMVS system gets the highest retrieval rate.

IV. CONCLUSION

In this paper, we propose a progressive transmission model for image descriptors in MVS based on the image saliency. Descriptors in different salient regions are transmitted to the server in order of their salience levels. In that case, we always achieve the right match with fewer descriptors rather than all the descriptors transmitted. Thus, the transmission latency can be decrease greatly. Meanwhile, without the noisy descriptors in the background region, higher retrieval rate are obtained. Moreover, on the server side, we propose a modified distance measure to further improve the matching accuracy. Extensive experiments on Stanford data set have shown that, the proposed MVSS system indeedly has lower transmission latency and higher retrieval rate compared with the classic MVS system.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation of China under grant nos.of 61173165,61370199,61300187, and the Fundamental Research Funds for the Central Universities under grant nos.of 2013QN044, DUT12RC(3)89,and DUT13JS04.

REFERENCES

[1] L.-Y. Duan, R. Ji, J. Chen, H. Yao, T. Huang, and W. Gao, "Learning from mobile contexts to minimize the mobile location search latency," *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 368–385, 2013.

[2] X. Shi, T. Sun, Y. Shen, K. Li, and W. Qu, "Tour-guide: providing location-based tourist information on mobile phones," in *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, pp. 2397–2401, IEEE, 2010.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[4] K. Li, H. Shen, F. Y. Chin, and W. Zhang, "Multimedia object placement for transparent data replication," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 18, no. 2, pp. 212–224, 2007.

[5] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, "Mobile visual search," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 61–76, 2011.

[6] J. Xia, K. Gao, D. Zhang, and Z. Mao, "Geometric context-preserving progressive transmission in mobile visual search," in *Proceedings of the 20th ACM international conference on Multimedia*, pp. 953–956, ACM, 2012.

[7] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, pp. 253–262, ACM, 2004.

[8] G. Shakhnarovich, *Learning task-specific similarity*. PhD thesis, Massachusetts Institute of Technology, 2005.

[9] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, pp. II–506, IEEE, 2004.

[10] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006*, pp. 404–417, Springer, 2006.

[11] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005.

[12] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod, "Chog: Compressed histogram of gradients a low bit-rate feature descriptor," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2504–2511, IEEE, 2009.

[13] V. R. Chandrasekhar, S. S. Tsai, G. Takacs, D. M. Chen, N.-M. Cheung, Y. Reznik, R. Vedantham, R. Grzeszczuk, and B. Girod, "Low latency image retrieval with progressive transmission of chog descriptors," in *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*, pp. 41–46, ACM, 2010.

[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1597–1604, IEEE, 2009.

[16] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, (Washington, DC, USA), pp. 1470–, IEEE Computer Society, 2003.

[17] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," in *Proceedings of the international conference on Multimedia*, MM '10, (New York, NY, USA), pp. 1469–1472, ACM, 2010.

[18] "Spectral / phase-based visual saliency." https://www.mathworks.com/matlabcentral/fileexchange/32455-spectral-phase-based-visual-saliency.

[19] M. Aly, M. Munich, and P. Perona, "Indexing in large scale image collections: Scaling properties and benchmark," in *Proceedings of the 2011 IEEE Workshop on Applications of Computer Vision (WACV)*, WACV '11, (Washington, DC, USA), pp. 418–425, IEEE Computer Society, 2011.

[20] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 965–968, IEEE, 2008.