J. of Mult.-Valued Logic & Soft Computing, Vol. 0, pp. 1–21
Reprints available directly from the publisher
Photocopying permitted by license only

A Low Overhead Progressive Transmission for Visual Descriptor Based on Image Saliency

YEGANG DU[1], ZHIYANG LI[1,*], MILOS STOJMENOVIC[2], WENYU QU[1] AND HENG QI[3]

[1]*College of Information Science and Technology, Dalian Maritime University, CN*
[2]*Department of Informatics and Computing, Singidunum University, RS*
[3]*School of Computer Science and Technology, Dalian University of Technology, CN*

Received: September 11, 2013. Accepted: May 19, 2014.

A typical mobile visual search (MVS) application generally follows the client-server architecture. Query images or these descriptors are transmitted from the mobile client to the remote server via the wireless network, to retrieval similar images from the database maintained in the server. Due to bandwidth constrained wireless networks, transmission latency is a bottleneck in present MVS. In some recent works, progressive transmission strategies have been proposed to reduce the transmission latency. The two main concerns in the progressive transmission are finding a proper priority of transmission and making up the recognition rate caused by the transmission loss. To address the two issues, a novel MVS framework is proposed in this paper, consisting of two main parts: a new progressive transmission model based on image saliency (MVSS) and a new matching metric designing for matching silent parts with whole images. Many experiments have been done on the public Stanford image set to evaluate the proposed MVSS system, and the results demonstrate that our framework not only reduces the transmission latency but also achieves a better retrieval accuracy, when comparing with the existing progressive transmission mechanisms.

Keywords: Saliency; bag of words; distance algorithm; mobile visual search; image retrieval

1 INTRODUCTION

In recent years, mobile terminals including smart phones and tablet PCs have significantly improved in the areas of communications and multimedia

* E-mail: lizy0205@dlmu.edu.cn

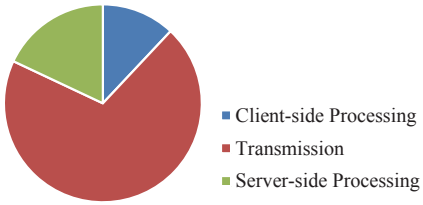


FIGURE 1
Component of latency.

[12, 21]. These devices are equipped with high-resolution cameras, touch screens and high-performance CPUs, which provide a perfect platform for Mobile Visual Search (MVS). MVS applications typically follow a client-server architecture. On the client side, a query photo taken by a user 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. Although some applications have been in business for a while, such as Google Goggles [1], Kooaba [2] and Layar [3], there are still some challenges performing mobile visual search. One of the biggest challenges is how to decrease the responding latency in both industry and academia [14].

Generally, there are three parts of the responding latency: client-side processing, transmission and server-side processing [25]. Because the data is transmitted via a bandwidth constrained wireless link, the transmission becomes the most time consuming part as depicted in Figure 1. That is to say, the performance of an MVS system usually heavily depends on how much data will be transmitted over the network. Many methods have been proposed in literature, concentrating on how to reduce the size of transmitted data. It has been shown that sending feature vectors or descriptors instead of original images is more efficient [14]. Thus, typical MVS system flow could be summarized as shown in Figure 2. Various presented image descriptors can be used in the feature extraction process.

The state-of-art visual descriptor is scale invariant feature transform (SIFT), which is also the most matured descriptor. However, the size of a SIFT descriptor is larger than other descriptors, even larger than the captured image sometimes, which makes it inappropriate to be directly used in MVS systems. Thus, many subsequent publications concentrate on how to reduce the 128-dimension SIFT descriptor into lower dimensional spaces, such as Locality Sensitive Hashing (LSH) SIFT [11], Similarity Sensitive Coding (SSC) SIFT [20] and Principal Component Analysis (PCA) SIFT [16]. There are also some other low-bit rate descriptors such as: Speeded-Up

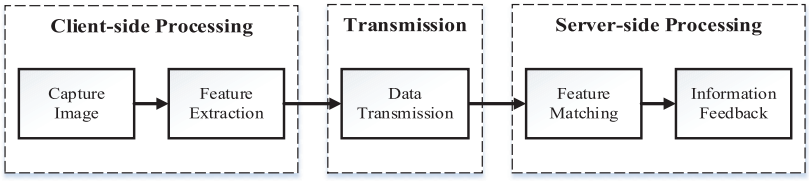


FIGURE 2
MVS flow.

Robust Features (SURF) [7], Gradient Location and Orientation Histogram (GLOH) [17] and Compressed Histogram of Gradients (CHoG) [8].

Another direction to decrease the responding latency is to improve the conventional pipeline by using a progressive transmission strategy. The method proposed by Chandrasekhar [9] extracts CHoG features, and then sorts the descriptors by Hessian responses. Those descriptors with higher Hessian response have higher priority, which means they will be transmitted first. The server begins to match the features when the first several descriptors arrive. However, this method does not take the geometric information in the captured image into consideration. In [25], a progressive transmission method is proposed based on geometric context preserving. Using this method, the captured image is divided into several blocks and features in the same block are transmitted as a unit. Compared to the previous transmitting strategy [9], 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 does not conform to the mechanism of human visual systems.

Given an image, human visual system tends to find the most informative regions and analyses the content of these regions, which are usually called salient regions in computer vision. Motivated by this observation, a progressive transmission strategy is presented in this paper based on image saliency. Upon capturing the image, we compute the saliency value of each pixel by the spectral residual approach [15] and divide the pixels into different saliency levels. After that, the feature points on the image are detected and these descriptive feature vectors are computed and then transmitted to the server. To decrease the responding latency, the computing and transmitting processes proceed in order of the saliency levels of the feature points, and this process will stop if an appropriate result is retrieved. SIFT descriptors are used in the experiment since SIFT is the most matured and its implementation is free on the Internet. The other state-of-art descriptors such as SURF and CHoG are also tested in the experiments.

Our main contributions can be summarized in three parts:

- Compared to the previous progressive transmission methods [9, 25], our presented strategy follows the mechanism of the human visual system, which can preserve the coherent regions in the image as a whole and effectively reduces the responding latency.
- In the classic Bag of Words (BoW) framework, images are encoded into global histograms to perform matching. Towards matching the transmitted silent regions and images in the database, we present a new distance measure, considering different magnitudes between these two kinds of histograms, which achieves a higher matching rate.
- Extensive experiments have been done on the public Stanford image date set, demonstrating that the proposed MVSS system generally outperforms the baseline method (sending all the feature vectors) and other progressive transmission methods in terms of recognition rate and responding latency.

2 DESCRIPTOR TRANSMISSION MODEL BASED ON SALIENCY

As explained in [25], points in the same object usually have similar feature information. The block-by-block transmission is an effective method to reduce the size of data transmitted, but also keeps geometric context information [14]. However, some mistakes can be made by roughly using grid lines to divide the captured image as shown in Figure 3. Points in the same object are divided into three blocks, which will break the geometric context.

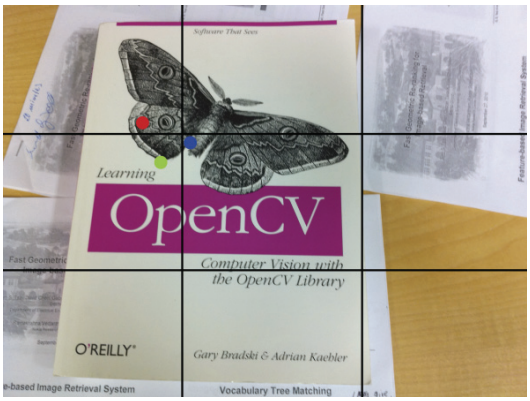
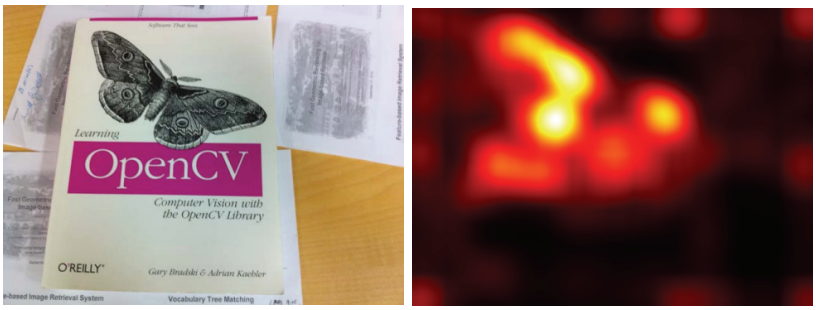


FIGURE 3
The mistakes caused by division.



(a) (b)

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

According to the characteristics of the human visual system, we propose a more reasonable transmission model based on image saliency.

Actually, various kinds of saliency detection methods have been presented in the literature [5], which are related to many applications, such as object recognition, automatic image cropping and image/video compressing. In this paper, we choose the saliency detection based on spectral residual (SR) presented in [15], because it can be implemented on a mobile terminal rapidly and easily. Given an image, its spectral residual(SR) is defined as the difference between the log spectrum of the image and the averaged log spectrum of sample natural images. The saliency map of the original image can be obtained by transforming spectral residual into spatial domain as shown in Figure 4, which is then 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 sections.

2.1 Client-side

Upon obtaining the captured image, the saliency values are computed by the mentioned SR approach for each image pixel. That is to say, a saliency matrix with the same size as the captured image has been obtained. Then M levels of salient regions which look like a terrace are determined by selecting different quantiles as shown in Figure 5. For example, we set level 4, which represents the top 20% salient region, and level 3 represents the top 40% salient region, then the rest can be done in the same manner. Particularly, level 0 represents the original image. Then from the figure below, we can see that different levels generally mean different significances. The higher levels we select, the more salient regions we will get. Content in the several most salient regions reflect the most significant sensory information of the image.

FIGURE 5
Different levels of salient regions.

The above saliency detection process is unsupervised. To improve the accuracy of the detection result, the salient regions can also be further delineated by mobile users themselves. This operation is an optional process, and can be easily realized on touch screens. Regions selected by the user are the ones they really care about and want to retrieve from the server. Consequently, the selected regions will be superimposed and gave higher saliency values.

Upon obtaining the saliency values of the query image as described above, the next step is computing image descriptors. 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 a high-dimensional descriptor, only the descriptors in salient regions are transmitted in order to reduce the transmission latency. SIFT computation consists of two steps: feature points detection and feature vector computation. As the first step is fast 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 feature

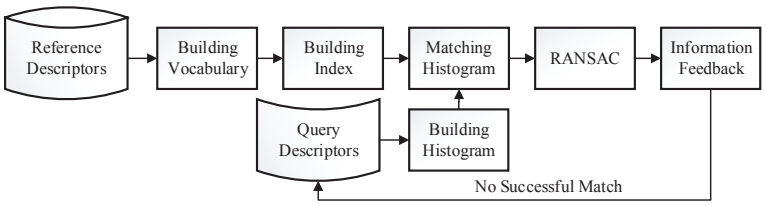


FIGURE 6
The bag of words framework.

vectors for points with higher saliency values are computed before the ones with lower values.

Features are then compressed and transmitted to the server at once. Features of the points belonging to the same saliency level are transmitted in one group. The server starts or restarts the retrieval process when it receives a complete group of features. There is usually no need to transmit all the features before the right feature matching is achieved. So our transmission mechanism can greatly reduce the amount of transmitted data, and then decrease the transmission latency.

2.2 Server-side

Due to the high performance of the server-side, we choose the BoW framework [22] to perform the feature matching task. 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 Figure 6.

At first, SIFT feature vectors are extracted for all the reference images in the database and constitute the training set. Then, a vocabulary or codebook is constructed by clustering the training vectors into k words. This process can be implemented by many quantization approaches, such as K-means [13], Approximate K-means (AKM) [19], Hierarchical K-means (HKM) [18]. AKM is chosen to be used in the experiment since it can handle large data sets. The size of codebook k is generally set based on the size of the training vectors. With a reasonable k, the result can be both accurate and efficient. It is noted that smaller codebooks always have less distinguishing ability but the larger ones will contain noise words.

The second step is building the index. Inspired by text retrieval, the reference images can also be inverted by their visual words. Thus, when retrieving images, a query image q only needs to match with the reference images inverted by the words occurring 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 the 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 first. Term frequency-inverse document frequency (tf-idf) normalization is a state-of-the-art method. Specifically speaking, for each word k_i in the vocabulary, its weighted word frequency is computed as followed:

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

A detailed discussion of these parameters is included in [22]. 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 is 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 Equation 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 the 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 the reference histogram in the database. Set I is the index set of visual words occurring in q, which obviously is a subset of the index set of the codebook. From Equation 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 experimental results in Section 3 demonstrate that the new distance measure outperforms other existing measures in most cases, and is more suitable to match histograms with different magnitudes.

Finally, the reference images are ranked by distances between them and the query image via Equation 2. Generally, in order to improve the retrieval rate, some candidate images are chosen and processed with the spatial checking process. RANSAC [10] 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 the most similar candidate images to the query image, and and provide this feedback to the users.

Note that when the first group of feature descriptors reaches the server, the server-side processing begins. As shown in Figure 6, the processing does not stop until the server finds the right match or no match at all after the last group of features is retrieved. Since the right match will mostly be obtained in the early groups containing more salient image information, the proposed MVSS system can reduce retrieval time and transmission latency greatly.

3 EXPERIMENT AND PERFORMANCE ANALYSIS

Due to the unstable network bandwidth in a realistic environment, we choose to do the MVS experiments on a PC. The experiments are carried out on a computer with a Intel(R) Core(TM) i3-2130 3.4GHz CPU and 2GB of 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 [23], Spectral/Phase-based Visual Saliency toolbox [4] and Caltech's Large Scale Image Search toolbox [6] respectively, which are all publicly available in the Internet.

Scale and constitution: The 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 on a 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 captured by a 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

FIGURE 7
Sample images in the Stanford data set. Note that the first row contains the reference images, while the last four rows are the query images. All eight categories are presented here.

(width 640 pixels), keeping the original image aspect ratio. All the parameters used in SIFT feature point detection and feature vector computation are chosen to be the default options.

When encoding the local feature vectors to the vocabulary by the AKM method, the vocabulary size k is an important parameter. It is pointed out 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 the 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.

3.1 The performance analysis of the vocabulary size *k*

The vocabulary size k will directly affect the effectiveness of the 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 the matching process will grow too.

To select a proper k for the 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. The average retrieval rate of all the eight categories are shown in Figure 8. From the results, we can see that even when a vocabulary size of k at $60K$ is selected, the recognition rate still appears to be increasing, which suggests that further gains could be achieved by increasing k.

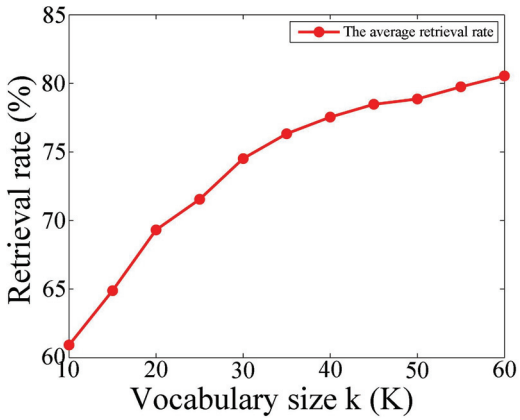


FIGURE 8
The chart of average result.

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 a proper choice considering both the efficiency and accuracy. So we set $k = 30K$ as the default value in the experiment.

3.2 Relation between the saliency level and its transmitted data size

In this experiment, we extract SIFT descriptors from all query images and transmit them in the order of their saliency levels. That is to say, the selected SIFT descriptors with higher saliency levels will be delivered by the network earlier. As the transmission latency heavily depends on the amount of SIFT descriptors transmitted over the network, 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 the 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 the Stanford data set, the transmission ratio of the descriptors in each saliency level M_i is computed by Equation 3

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

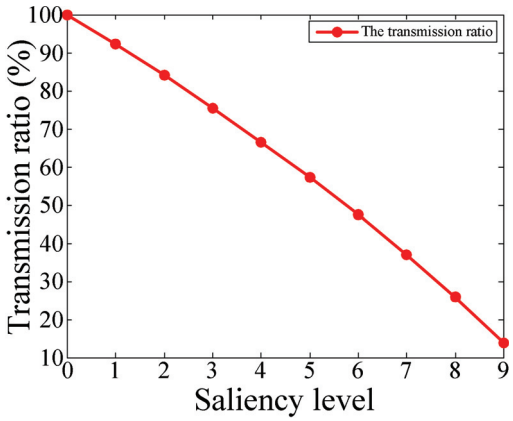


FIGURE 9
The mean value of the saliency level and corresponding R_i.

The average transmission ratio in each saliency level among all the images are illustrated in Figure 9, which shows that the relationship between saliency level and corresponding R_i is approximately linear. That is, when the saliency level turns higher, the corresponding R_i will low down with similar portion.

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 through the network. For instance, if the server finds the right match with level 9, our method can achieve nearly 90 percent reduction on transmission latency.

3.3 Analysis on retrieval rate and the saliency level

In this section, we compare our method with the baseline method (SIFT + BOW) in Figure 10 and Figure 12. Although the saliency level is chosen as 5 in the MVSS system, we set the 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 horizontal ordinate represents saliency level M which means transmitting the SIFT descriptors in the first M saliency levels to do the retrieval experiment. The ordinate represents the corresponding retrieval rate.

From Figure 10, we can see that our method performs better than the baseline method on seven categories of data sets. 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 the “saliency reversal” problem in [24].

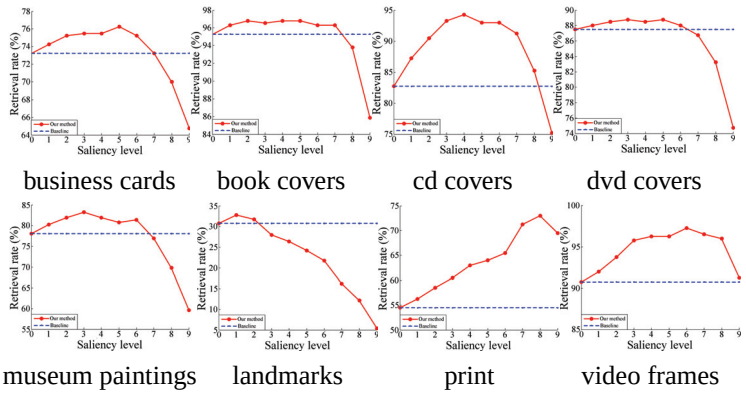


FIGURE 10
The contrast between our method and baseline method.

Saliency reversal is caused by mistakes that regard the background as the foreground. The images in Figure 11 selected from the category of landmarks, are instances of "saliency reversal". Cars are the main body of the selected regions in the first three salience levels. However, the actual retrieval object is the building. We will optimize our method to overcome this problem in our future research.

Figure12 shows the average result on the Stanford data set. We can see that the retrieval rate of our method first increases and then decreases, with the increase of the salience 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 decrease as well. Then fewer descriptors

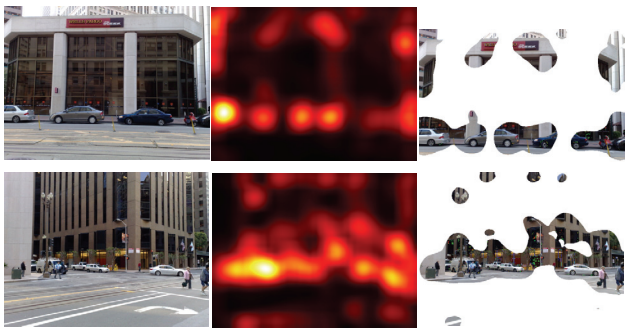


FIGURE 11
The "saliency reversal" problem. Take saliency level 7 for example. These two pictures are from landmarks on Stanford.

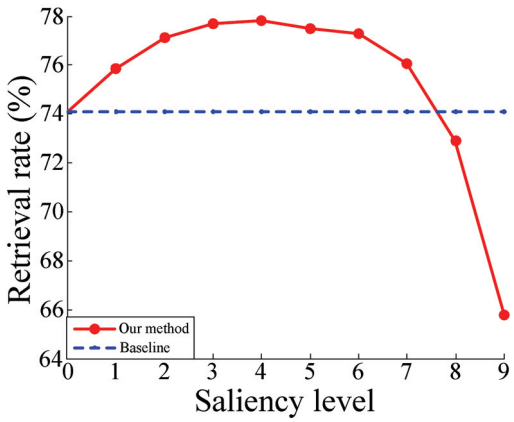


FIGURE 12
The average result.

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

3.4 Analysis on the proposed distance measure

To further discuss the performance of our distance measure defined in Equation 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 cases. 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 retrieval rates and the average retrieval rates on the eight categories utilizing the different distance algorithms are shown in Figure 13 and Figure 14, respectively.

From the bar graph, we can see that our distance measure achieves the highest retrieval rates when the salience level M is higher than 2, on the contrary, achieves lower retrieval rates than L_1 distance when M is lower than 2. This means that when only a small part of all descriptors is encoded,

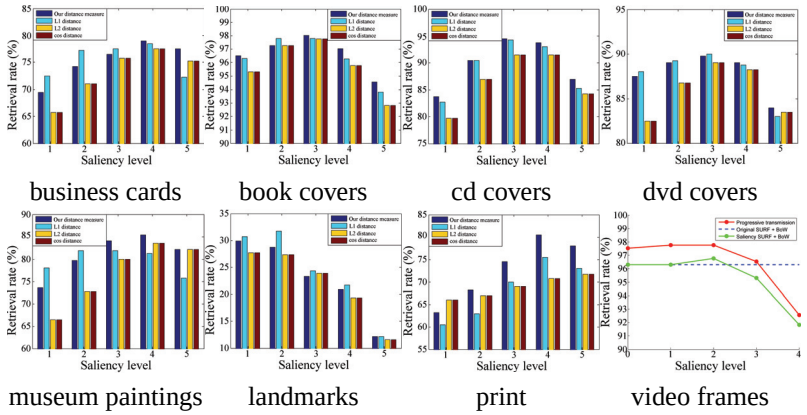


FIGURE 13
The contrast between our distance measure and other distance measures.

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

3.5 Experiments with CHoG and SURF

In this part, we further verify the effectiveness of our optimized transmission model based on saliency with different visual descriptors. As described in

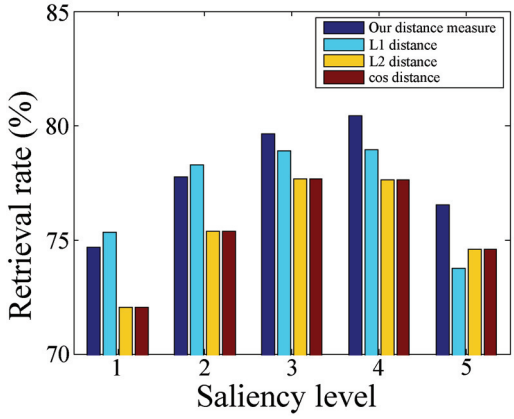


FIGURE 14
The average result.

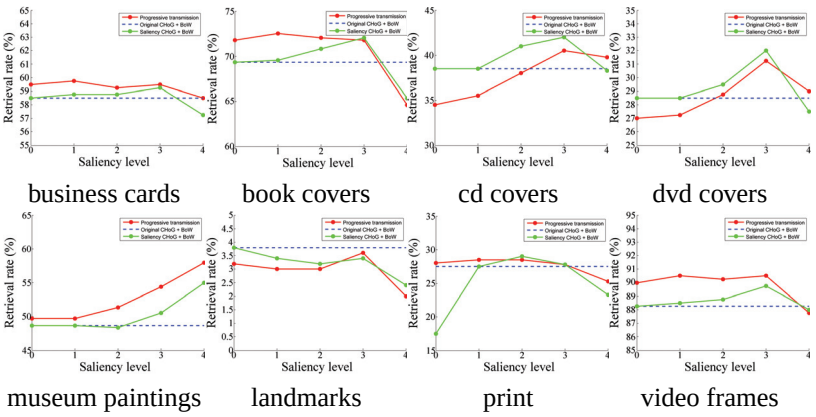


FIGURE 15
The retrieval rate on Stanford dataset. The parameters are equal to those of the previous experiment, except that the descriptor is changed to CHoG.

Section 1, there are some kinds of low-bit rate descriptors in the field of image retrieval research. In this section, we choose two of the most representative low-bit rate descriptors to serve as contrasts to the state-of-art visual descriptor SIFT. CHoG is proposed by Chandrasekhar [8]. They first exploit the underlying gradient statistics, then they use tree coding techniques to quantize the histograms into low bit-rate feature descriptors. SURF by Bay [7] works by way of relying on integral images for image convolutions and building on the strengths of the leading existing detectors and descriptors.

In order to do a fair experiment, all parameters are set to the same as in the baseline method. In the descriptor extraction step, we extract the CHoG descriptors and SURF descriptors utilizing the existing toolbox provided by the authors. As in the previous experiments, descriptors in the higher saliency region are transmitted preferentially. And in the histogram distance calculation step, our modified distance measure is also employed to match the transmission strategy based on saliency. The results of the two experiments are represented respectively by Figure 15 and Figure 16.

One can see that our optimized transmission model can improve the retrieval rate on experiments both with CHoG and SURF. Although our transmission strategy does not get the expected results on some individual categories, the average result can still reach a satisfactory level. Figure 17 shows the average retrieval rates of CHoG and SURF, from which we can notice that our transmission strategy improves the average retrieval rate by about 2%. This experiment illustrates that our progressive transmission model is a universally applicable method which can be widely used with all kinds of visual descriptors.

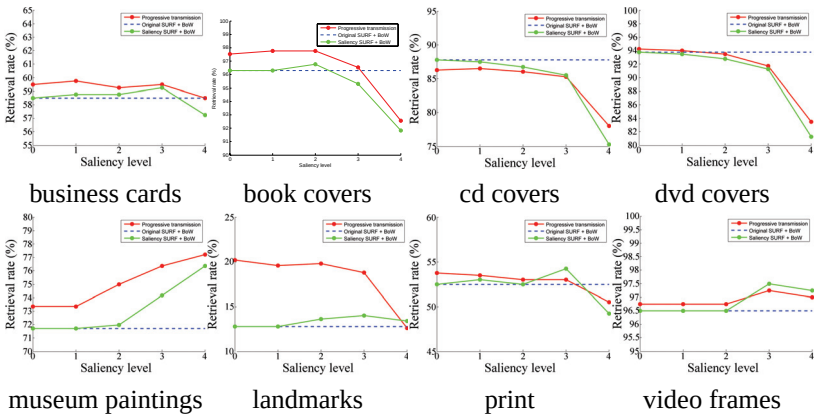


FIGURE 16
The retrieval rate on the Stanford dataset. The parameters are equal to those of the previous experiment, except that the descriptor is changed to SURF.

3.6 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 since it outperforms other distance measures in the 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

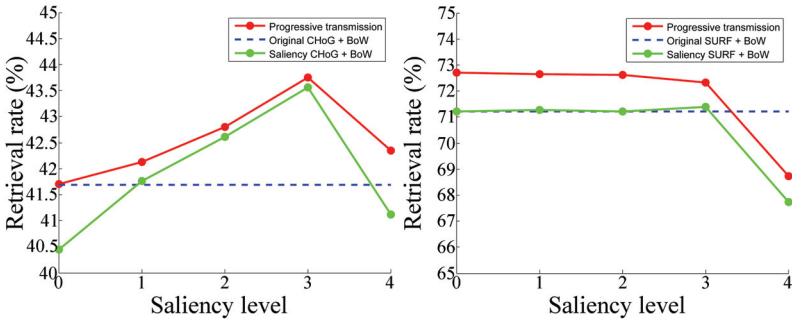


FIGURE 17
The average retrieval rate of two widely used descriptors. (a) shows the average retrieval rate of CHoG and (b) shows the average retrieval rate of SURF.

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 1
MVSS system.

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 2
Original 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%

TABLE 3
Saliency SIFT + BoW.

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 1 2 3 and Figure 18.

As depicted in Table 1, the retrieval rate of our method further improves 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

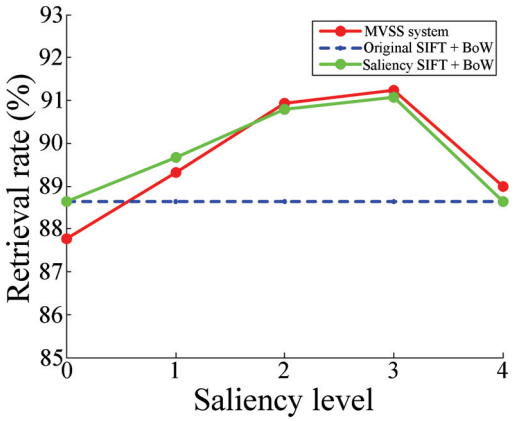


FIGURE 18
The average result.

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.

Experiments have shown that our method can highly improve the matching rate with fewer feature descriptors on those kinds of images with salient objects. Our method can indeed decrease the transmission latency and increase the retrieval rate at the same time. These distinguishing characteristics and advantages make our method very suitable for retrievals of clear targets in MVS.

4 CONCLUSION

In this paper, we propose a progressive transmission model for image descriptors in MVS based on image saliency. Descriptors in different salient regions are transmitted to the server in the 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 rates are obtained. Moreover, on the server side, to match the two kinds of histogram encodings by the transmitted salient regions and images, we present a modified matching measure which further increases the retrieval rate, Extensive experiments on the Stanford data set have shown that, the proposed MVSS system indeed has lower transmission latency and higher

retrieval rate compared with the classic MVS system. In future research, we will focus on designing light descriptors for MVS. Furthermore, since the Stanford data set is a relatively small one, we will prepare to do experiments on some larger data sets in the future.

ACKNOWLEDGMENTS

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. This work was partially supported by the following grant: "Digital signal processing, and the synthesis of an information security system", TR32054, Serbian Ministry of Science and Education.

REFERENCES

[1] Google goggles. http://www.google.com/mobile/goggles.

[2] Kooaba. http://www.kooaba.com.

[3] Layar. http://www.layar.com.

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

[5] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. (2009). Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597–1604. IEEE.

[6] Mohamed Aly, Mario Munich, and Pietro Perona. (2011). Indexing in large scale image collections: Scaling properties and benchmark. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 418–425. IEEE.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. (2006). Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer.

[8] Vijay Chandrasekhar, Gabriel Takacs, David Chen, Sam Tsai, Radek Grzeszczuk, and Bernd Girod. (2009). Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2504–2511. IEEE.

[9] Vijay R Chandrasekhar, Sam S Tsai, Gabriel Takacs, David M Chen, Ngai-Man Cheung, Yuriy Reznik, Ramakrishna Vedantham, Radek Grzeszczuk, and Bernd Girod. (2010). Low latency image retrieval with progressive transmission of chog descriptors. In *Proceedings of the 2010 ACM multimedia workshop on Mobile cloud media computing*, pages 41–46. ACM.

[10] Sunglok Choi, Taemin Kim, and Wonpil Yu. (1997). Performance evaluation of ransac family. *Journal of Computer Vision*, 24(3):271–300.

[11] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni. (2004). Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 253–262. ACM.

[12] Ling-Yu Duan, Rongrong Ji, Jie Chen, Hongxun Yao, Tiejun Huang, and Wen Gao. (2013). Learning from mobile contexts to minimize the mobile location search latency. *Signal Processing: Image Communication*, 28(4):368–385.

[13] Charles Elkan. (2003). Using the triangle inequality to accelerate k-means. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, volume 20, page 147.

[14] Bernd Girod, Vijay Chandrasekhar, David M Chen, Ngai-Man Cheung, Radek Grzeszczuk, Yuriy Reznik, Gabriel Takacs, Sam S Tsai, and Ramakrishna Vedantham. (2011). Mobile visual search. *Signal Processing Magazine, IEEE*, 28(4):61–76.

[15] Xiaodi Hou and Liqing Zhang. (2007). Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

[16] Yan Ke and Rahul Sukthankar. (2004). 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*, volume 2, pages 11–506. IEEE.

[17] Krystian Mikolajczyk and Cordelia Schmid. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630.

[18] David Nister and Henrik Stewenius. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2161–2168. IEEE.

[19] Dan Pelleg and Andrew Moore. (1999). Accelerating exact k-means algorithms with geometric reasoning. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 277–281. ACM.

[20] Gregory Shakhnarovich. (2005). *Learning task-specific similarity*. PhD thesis, Massachusetts Institute of Technology.

[21] Xiaoyu Shi, Ting Sun, Yanming Shen, Keqiu Li, and Wenyu Qu. (2010). Tour-guide: providing location-based tourist information on mobile phones. In *Computer and Information Technology (CIT), 2010 IEEE 10th International Conference on*, pages 2397–2401. IEEE.

[22] Josef Sivic and Andrew Zisserman. (2003). 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, pages 1470, Washington, DC, USA. IEEE Computer Society.

[23] Andrea Vedaldi and Brian Fulkerson. (2010). Vlfeat: an open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, MM '10, pages 1469–1472, New York, NY, USA. ACM.

[24] Zheshen Wang and Baoxin Li. (2008). A two-stage approach to saliency detection in images. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 965–968. IEEE.

[25] Junhai Xia, Ke Gao, Dongming Zhang, and Zhendong Mao. (2012). Geometric context-preserving progressive transmission in mobile visual search. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 953–956. ACM.