Progressive transmission based on wavelet used in mobile visual search

Siyang Miao

College of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning Province, P.R.China
Fax: +86-0411-84723505 E-mail: italymm0412@gmail.com

Zhiyang Li

College of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning Province, P.R.China
Fax: +86-0411-84723505 E-mail: lizy0205@dlmu.edu.cn

Wenyu Qu

College of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning Province, P.R.China
Fax: +86-0411-84723505 E-mail: wenyu@dlmu.edu.cn

Yegang Du

College of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning Province, P.R.China

Songhe Wang

College of Information Science and Technology,
Dalian Maritime University, Dalian, Liaoning Province, P.R.China

Heng Qi

School of Computer Science and Technology,
Dalian University of Technology, Dalian, Liaoning Province, P.R.China

Abstract: Mobile visual search (MVS) has become a hot topic in both academia and business circles. A typical MVS system generally uses the camera phone to initiate search queries, following a client-server architecture. Transmission overload turns out to be a bottleneck due to the bandwidth constrained wireless link. An effective way to reduce transmission overload is exploiting progressive transmission strategy, which probably obtains the correct match via a few more important image descriptors delivered. However, the acceleration effects of existing methods are often limited, because they only decrease the amount of descriptors delivered, and important descriptors with high dimension still need being entirely delivered. To compress the descriptors and further reduce network latency, many methods have been offered, such as PCA-SIFT, SURF, CHoG. One of the problem is that they can not restore the original descriptor and generally the more compressive the descriptor is, the lower the precision is. To conquer the above problems, we propose a progressive transmission strategy based on multi-level wavelet decomposition and reconstruction theorem, which can both compress the descriptor and restore the original data at the same time. Extensive experiments have been done on the public Stanford MVS database, demonstrating that our proposed progressive transmission strategy outperforms other strategies based on SIFT and SURF descriptors directly, when delivering the same amount of data.

Keywords: Progressive transmission; wavelet theorem; mobile visual search.

Reference : Siyang Miao, Zhiyang Li, Wenyu Qu, Yegang Du, Songhe Wang and Heng Qi. (2013) 'Progressive transmission based on wavelet used in mobile visual search', *International Journal of Embedded Systems*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Siyang Miao is currently completing a MS in College of Information Science and Technology at the University of Dalian Maritime, China. Her main research interests include Information Retrieval and Computer Vision.

Zhiyang Li is an Assistant Professor in College of Information Science and Technology, Dalian Maritime University, China. He received his Ph.D degree from Dalian University of Technology, China, in 2011. His research areas include Information Retrieval and Computer Vision.

Wenyu Qu is a Professor at the College of Information Science and Technology, Dalian Maritime University, China. She got her doctor degree from Japan Advanced Institute of Science and Technology in 2006. Her research interests include Information Retrieval, Cloud Computing.

1 Introduction

With the popularization of high-resolution camera embedded mobile devices, mobile phones have evolved into powerful image video processing devices, equipped with color displays, and hardware-accelerated graphics. They are also equipped with location sensors, GPS receivers, and connected to broadband wireless networks allowing fast transmission of information. Many new researches have been working on smart phones Leung, CHC and Leung, TLY. (2008); Li, Keqiu (2011). This enlightens a new class of applications which use the camera phone to initiate search queries about objects in visual proximity to users. Such applications can be used for identifying products, comparison shopping, finding information about movies, CDs, real estate or products of the visual arts Chandrasekhar et al. (2010). First commercial applications of such systems include Google Goggles(Google, 2009), Nokia Point and Find(Nokia, 2006), Kooaba(Kooaba, 2009), and Snaptell(Amazon, 2007).

To this end, many of these applications follow a client-server architecture. On the client side, the mobile user takes a query photo, which is transmitted to the remote server to identify its corresponding content through visual matching. Then, the remote server maintaining a sizable database returns the searching results. In many cases, the query photo or its descriptor is delivered over a bandwidth constrained wireless link. Thus, the user experience heavily depends on how much data to transmit. And transmission overload turns out to be a bottleneck in most existing mobile visual search applications. To reduce the amount of the data transmitted, and thus reduce network latency, many methods are proposed, which are concentrated on the server side Li et al (2007) or client side.

One direct way to reduce transmission latency is compressing the image descriptors or designing lightweight descriptors. Since Lowe's paper in 1999 Lowe, David G. (1999), the highly discriminative SIFT descriptor remains the most popular descriptor. But, as a 128-dimension descriptor, SIFT is conventionally stored as 1024 bits(8 bits/dimension). This makes the size of SIFT descriptor from an image is typically larger, making it unsuitable for mobile applications. Several compression schemes have been proposed which can be broadly categorized into schemes based on hashing, transform coding and vector quantization Chandrasekhar et al. (2012). Representative methods of these schemes include PCA-SIFT Ke et al. (2004), SURF Bay et al (2008), CLOH Mikolajczyk et al (2005) and CHoG Chandrasekhar et al. (2012). These methods all compress SIFT descriptor in someway, but they also have their own defects. They cannot restore the original descriptors and the more compressive the descriptor is, the lower the precision is. In an extreme case, there would be no matching image returned back to the client side, when the compressed descriptor contains too little useful information to search in a giant database.

Progressive transmission is another effective way of reducing transmission latency in mobile visual search Chandrasekhar et al. (2010) Xia et al. (2012). They firstly extract image features on the client side, then simply sort the query descriptors by their importance. Descriptors with more importance are transmitted to server prior to others. On the server side, matching process is started as soon as the server receives a group of descriptors, and stops data transmission when the query image is matched correctly. Under sending features mode, progressive transmission method is not only effective to decrease the network latency because some overlapped time between network and server is saved, but also reduces the data size sent over the network for only part of the query features is needed before the right match is found. However, the acceleration effect of existing progressive transmission strategies are often limited because the descriptors sorted at the front order still need being entirely delivered, which are generally high-dimension vectors.

To overcome the above problems, in this article we propose a progressive transmission strategy as shown in Fig. 1 based on multi-level wavelet decomposition and reconstruction theorem, which enables efficient mobile visual search even in a bandwidth constrained wireless link with two major contributions:

Firstly, our method utilizes multi-level wavelet decomposition theorem to reduce the dimension of the SIFT descriptor in order to decrease the data transmitted over a bandwidth constrained wireless link. On the server side, a multi-level bag-of-words (BOW) model is built, of which each level corresponds to the same level descriptor, making searching more rapidly and more accurately. Note that, it is somehow time-consuming to build a multi-level BOW model on the server side, but this work has been done offline, which means it is not time-consuming to the phone users.

Secondly, when the extreme case mentioned above happens, our method can restore the original image descriptors using wavelet reconstruction theorem to obtain matching images similar to the original image. This guarantees that the phone users can always get a correct matching result from the server side under any circumstance.To our knowledge, there is no other proposed strategy that uses both multi-level wavelet decomposition and reconstruction theorem on mobile visual search(MVS), which is a technical novelty of our paper.

The remainder of this paper is organized as follows. Section 2 provides a brief description of our method. Section 3 details our progressive transmission based on wavelet used in MVS. Section 4 exhibits and analyzes our experiments. Section 5 summarizes the contributions of this paper.

2 Overview

Our proposed progressive transmission method based on wavelet follows the traditional client-server architecture. Fig. 1 gives a high-level overview of our progressive transmission method. Due to the progressive transmission, we do not have to transmit all the data. In Fig. 1 full lines stand for fixed part of the transmission and dotted lines represent auxiliary part. Each stage will be elaborated in the following Section 3.

To introduce wavelet theorem into SIFT descriptor, we refer to a 128-dimension SIFT descriptor S as a one-

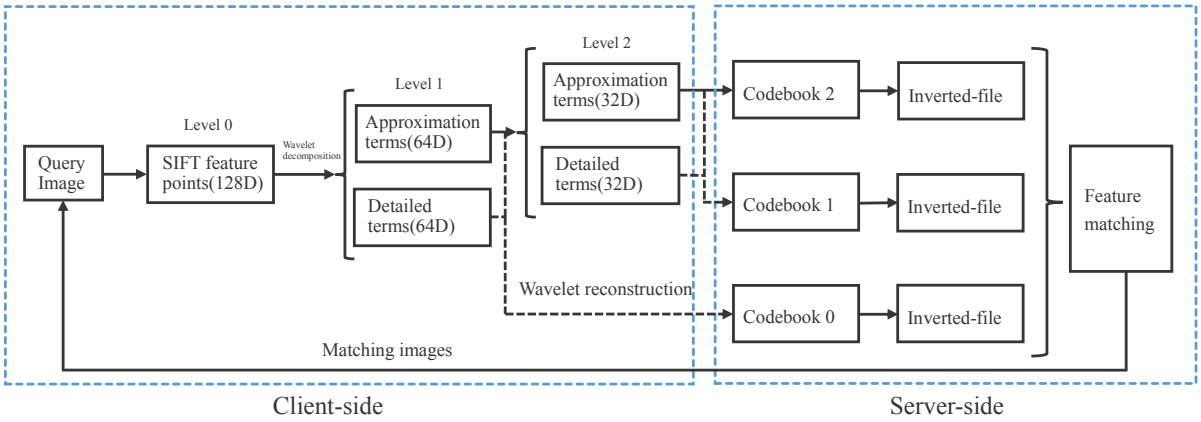


Figure 1 The overview of our progressive transmission based on wavelet used in mobile visual search.

dimension discrete signal $S(t)$ oscillating and decaying in the time domain. Since an image I is commonly represented by a collected set of local descriptors such as SIFT, I also can be represented by a set of one-dimension discrete signal with the same size. Here the size of the set is the number of feature points for the image.

In the client-side, for an image I taken by the mobile user, its SIFT descriptors are computed and then decomposed by multi-level wavelet. Haar wavelet is chosen as the default version. One of the reasons is that Haar wavelet is not only easiest but also liable to code implementation. Since the goal of our strategy is to reduce network latency and increase user experience on mobile devices, Haar wavelet is fully competent of this job. Another reason is the lightness of Haar wavelet, making it proper for mobile devices. Also, it ensures fast computational speed, which is essential for mobile devices. All the descriptors of I are processed as a whole to further improve the efficiency of the algorithm. Approximation coefficients represent the average part of the descriptors, which is usually assumed as an approximation version, are transmitted to the server to perform a retrieval task. The multi-level wavelet decomposition is discussed in Section 3.2 and 3.3.

In the server-side, BoW framework is utilized to accomplish the image retrieval. Classic BoW generally built one codebook for all the training images by quantization methods such as approximate K-means. For the multi-level decomposition and transmission characteristics of our method, a multi-codebook is built in the server-side. Each level has its own codebook which is obtained by clustering the approximation coefficients in this level. Once the approximation coefficients of the query image are transmitted into the server-side, they will be encoded into a global word representation according to the pre-built codebook of its level, and do feature matching process with the images inverted by the same word. If no right match is found, the approximation coefficients are combined with the subsequent detailed coefficients to reconstruct the former level approximation coefficients and treated as a new input of our proposed BoW framework. Details of the process in the serve side are explained in Section 3.4.

3 Methodology

3.1 The review of SIFT descriptor

Since Lowe's paper in 1999 Lowe, David G. (1999), the highly discriminative SIFT descriptor remains the most popular descriptor in computer vision. SIFT, as described in Lowe, David G. (1999), consists of two major stages: feature point detection and feature descriptor computation.

In the first stage, potential feature points are identified by scanning the image over the location and scale. This is implemented efficiently by constructing a Gaussian pyramid and searching for local peaks (termed key points) in a series of difference-of-Gaussian (DoG) images. Candidate key points are localized to sub-pixel accuracy.

The second stage builds a local image descriptor for each key point, based upon the image gradients in its local neighborhood. Specifically speaking, A 4×4 array of histograms, each with 8 orientation bins, captures the rough spatial structure of the patch around the key point. This 128-element vector is then normalized to unit length and thresholded to remove elements with small values.

The standard SIFT keypoint descriptor representation is remarkable in several respects: (1) the representation is carefully designed to avoid problems due to boundary effects-smooth changes in location, orientation and scale do not cause radical changes in the feature vector; (2) it is fairly compact, expressing the patch of pixels using a 128 element vector; (3) while not explicitly invariant to affine transformations, the representation is surprisingly resilient to deformations caused by perspective effects. These characteristics are evidenced in the excellent matching performance against competing algorithms Mikolajczyk et al (2005).

However, as a 128-dimension descriptor, SIFT is conventionally stored as 1024 bits (8 bits/dimension). Alas, the size of SIFT descriptor data from an image is typically larger than the size of the JPEG compressed image itself, making it unsuitable for mobile applications. Our initial goal is to compress the 128-dimension descriptor from SIFT, making it not only reasonable small for mobile visual

search, but also have satisfactory searching results. To achieve this, we exploit multi-level wavelet decomposition and reconstruction theorem to post-process the SIFT descriptor, and build a multi-level bag-of-word model on the server side. We will discuss these techniques detailedly in the succeeding parts.

3.2 Multi-level wavelet decomposition and reconstruction

As discussed above, the SIFT descriptor is too large to use for mobile visual search because of the bandwidth constrained wireless link. To reduce the network latency, a wavelet-based SIFT descriptor is designed and utilized for mobile visual search.

As mentioned, the SIFT descriptor S is generally represented by a 128-dimension vector, which can be assumed as a one-dimension discrete signal $S(t)$. An example for the SIFT descriptor is illustrated in Fig. 2(a). As a mathematical tool, discrete wavelet transform (DWT) is popular in analyzing signals, images and other kinds of data.

Common wavelet basis set begins with two orthogonal functions: the scaling function $\phi(t)$ satisfying $\int \phi(t) = A$ and the wavelet function $\psi(t)$ satisfying $\int \psi(t) = 0$. One can obtain a complete basis set $\{\phi_k(t), \psi_{jk}(t)\}$, by scaling and translating these two functions. Different versions of wavelet always have different basis sets. Using the mentioned basis set functions, the wavelet decomposition of a signal $S(t)$ can be written according to the following Eq. 1

$$S(t) = \sum_{k=-\infty}^{\infty} c_k \phi_k(t) + \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} d_{jk} \psi_{j,k}(t), \tag{1}$$

where the coefficients c_ks and $d_{j,k}$s are explicitly given by Eq. 2 according to the property of basis functions.

$$c_k = \int S(t)\phi_k(t)dt, \quad d_{jk} = \int S(t)\psi_{jk}(t)dt. \tag{2}$$

Coefficients defined in Eq. 2 are approximation and detail coefficients for signal $S(t)$. Approximation coefficients c_ks represent the average part of $S(t)$, which is an approximation version of the original signal. Detail coefficients d_{jk}s capture the variations of the signal at different scales. The two pictures in the second row of Fig. 2 illustrate the approximation and detail coefficients of the original signal (Fig. 2(a)) computed by Haar wavelet. From it, one can tell the 64 approximation coefficient is a good approximation of the original signal. Meanwhile, with the coefficients and basis set function, the signal can be reconstructed by Eq. 1.

It is obvious that one can further decompose the approximation coefficient to make a multi-level decomposition of the $S(t)$. Seeing the third row of Fig. 2. The wavelet transform enables to give two interesting applications to the SIFT descriptor: compressive and progressive transmission. The approximation terms can be assumed as a compressive version of original SIFT and transmitted firstly to the server. In the next section, the decomposition and progressive transmission of SIFT are discussed.

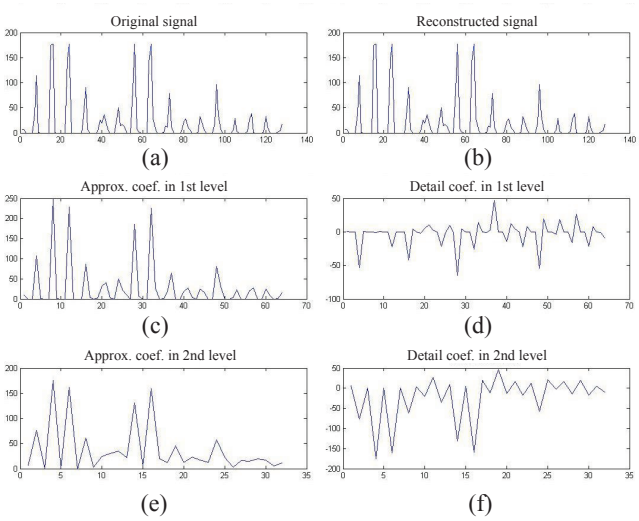


Figure 2 An example for the decomposition and reconstruction of a SIFT descriptor.

3.3 Client-side

For a query image I, taken by the user, the mobile computes its SIFT descriptors firstly. The SIFT descriptors can be represented by a feature matrix V_{mn}, having m rows and n columns. Here, m represents the number of feature points and generally $n = 128$. Upon obtaining the SIFT descriptors, the next thing is decomposing it by wavelet transform as discussed above. We make use of Haar wavelet for its simplicity and efficiency. The scaling function $\phi(t)$ and wavelet function $\psi(t)$ are chosen as following Eq. 3 and Eq. 4, respectively.

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\psi(t) = \begin{cases} 1 & 0 \le t < 1/2, \\ -1 & 1/2 \le t < 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

To further improve the efficiency of the algorithm, the mentioned feature matrix V is decomposed as a whole according to the Eq. 5.

$$V_{mn}^{(1)} = V_{mn} * A_{nn}^{(1)}, \tag{5}$$

$$A_{nn}^{(1)} = \begin{pmatrix} 1/2 & 0 & \cdots & 0 & 1/2 & 0 & \cdots & 0 \\ 1/2 & 0 & \cdots & 0 & -1/2 & 0 & \cdots & 0 \\ 0 & 1/2 & \cdots & 0 & 0 & 1/2 & \cdots & 0 \\ 0 & 1/2 & \cdots & 0 & 0 & -1/2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/2 & 0 & 0 & \cdots & 1/2 \\ 0 & 0 & \cdots & 1/2 & 0 & 0 & \cdots & -1/2 \end{pmatrix},$$

where $A_{nn}^{(1)}$ is the first level decomposition matrix. The first 64 columns of $V_{mn}^{(1)}$ are the approximation terms of original SIFT

descriptor and the next 64 columns are the corresponding detailed terms. To constitute a multi-level decomposition, $V_{mn}^{(1)}$ can be multiplied by the further level decomposition matrix according to the Eq. 6

$$V_{mn}^{(t)} = V_{mn}^{(t-1)} * A_{nn}^{(t)}, \tag{6}$$

$$A_{nn}^{(t)} = \begin{pmatrix} A_{n'n'}^{(1)} & 0 \\ 0 & E_{(n-n')(n-n')} \end{pmatrix}.$$

Where $A_{nn}^{(t)}$ is the t level decomposition matrix, $n' = n/(2^{t-1})$ and E is the unit matrix. Thus, the SIFT descriptor V is decomposed into a matrix $V_{mn}^{(t)}$ via t-level decomposition. It is obvious that the first $128/(2^t)$ columns of $V_{mn}^{(t)}$ are the approximation terms of the t-level decomposition. The next $128/(2^t)$ columns of $V_{mn}^{(t)}$ are detailed terms of the t-level decomposition, the next $128/(2^{(t-1)})$ columns of $V_{mn}^{(t)}$ are approximation terms of the $t-1$-level decomposition. Based on the reconstruction theory, the original SIFT descriptor V can be reconstructed by $V^{(t)}$ via Eq. (1). Towards Harr wavelet, a reconstruction of $V_{mn}^{(t-1)}$ via $V_{mn}^{(t)}$ can be achieved by the following Eq 7.

$$V_{mn}^{(t-1)} = V_{mn}^{(t)} * \mathbf{Inv}(A_{nn}^{(t)}). \tag{7}$$

Thus, one advantage of our transmission methods is that V_{mn} can be transmitted at various levels of details, starting with the approximation terms (the first $128/(2^t)$ columns), and over time more and more details are added in, until a satisfactory matching result is received from the server side. Note that, progressive transmission has one major advantage: the client side can halt the process and move on to something else if user decides, based on early feedback results, that satisfactory matching images have been received from the server side only with part of the SIFT descriptor transmitted Mulcahy, Colm (1997). This decreases the amount of data transmitted by bandwidth constrained wireless link, reducing waiting time and enhancing the user experience.

3.4 Server-side

The task of server-side is to perform feature matching between the transmitted image and images in the database maintained by the server. One widely used matching framework is BoW framework Sivic, Josef et al. (2003), which utilizes the inverted index method and retrieves large image databases rapidly and effectively. A classic BOW consists of four steps: codebook building via quantization methods such as approximate K-means, encoding the local features into a global word representation according the codebook, feature matching and spatial checking process via RANSAC.

A main difference between the proposed method and the classic BoW framework is the codebook building process. Since the image descriptor is delivered in a progressive manner, a multi-codebook is built here.

For each image in the database, its SIFT descriptor is decomposed by multi-level wavelet as mentioned above. In each level, a codebook is built by clustering the approximation terms of the SIFT for all the images in the database. Thus, if t-level wavelet decomposition is done, $t+1$ codebooks will be obtained. That is also to say, all the images are inverted by $t+1$ codebooks and encoding into $t+1$ global representations. Although these steps are somehow time-consuming, they only need to be carried out offline, which means the process would not produce any latency during retrieval.

Once the approximation terms $V_{mn}^{(t)}$ of the query image are transmitted to the server-side. $V_{mn}^{(t)}$ will be encoded into a global word representation according to t-level pre-built codebook, and do feature matching process with the images inverted by the same word. If no right match is found, $V_{mn}^{(t)}$ will be combined with its corresponding detailed term subsequently transmitted to reconstruct the upper level approximation terms $V_{mn}^{(t-1)}$. $V_{mn}^{(t-1)}$ will then do the same thing as $V_{mn}^{(t)}$. The client side can halt the matching process and transmission process once right match is found.

We point that another advantage of the wavelet-based image descriptor is its reconstruction function. As mentioned above, other compressed descriptors of SIFT, like PCA-SIFT, SURF and CHoG, unable to reconstruct the original SIFT descriptor. When the extreme case happened, that the compressed descriptor carries too little useful information to get a matching image, there will be no feedback result to users, which is absolutely unacceptable. Unlike these methods, our method enables to reconstruct the original SIFT descriptor by wavelet reconstruction theorem. Thus, this guarantees that the phone users can always get a matching result from the server side under any circumstance.

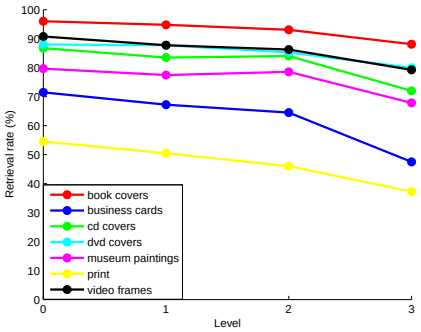
4 Experimental results

In this section, we evaluate the performance of the proposed system via image retrieval experiments on Stanford data set, which is carried out on a computer with a Core(TM) i3-2130 3.4GHz CPU and 2GB memory. The code is implemented in MATLAB 2012a with some parts written in C with a MEX interface. At the client side, the implementation of image descriptors SIFT and SURF are based on VLFeat toolbox Vedaldi, Andrea and Fulkerson, Brian (2010) and SURFmex toolbox. At the server side, the bag-of-word image retrieval framework is performed by using the Caltech Large Scale Image Search toolbox Aly et al. (2011).

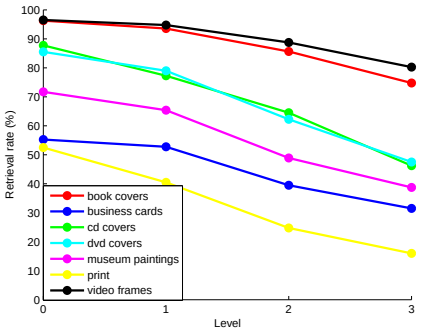
The Stanford data set consists of eight categories, including book covers, business cards, cd covers, dvd covers, landmarks, museum paintings, print and video frames. All the categories contain two kinds of images: query images and reference images. Query images are taken by different kinds of mobile phones. For the experiments, each reference image is regarded as standard, and the query images are retrieved to find these corresponding reference image. All the images in the database are converted to gray-scale format with a fixed size (width 480 pixels), keeping the original image aspect ratio.

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 server side maintains. The retrieval rates in each category are computed respectively.

Next, we describe the parameters used in the experiments. At the client side, the parameters used for SIFT and SURF are selected as the default case of the toolboxes. The default values of decomposition level for SIFT and SURF are 2 and 1. At the server side, one important parameter of BOW framework is the size of vocabulary k. Larger vocabulary tends to achieve higher retrieval rates. In the following, performance analysis of the mentioned parameters is discussed, then retrieval experiments on Stanford data set are done to compare our approach with other progressive transmission approaches Chandrasekhar et al. (2010).



(a)



(b)

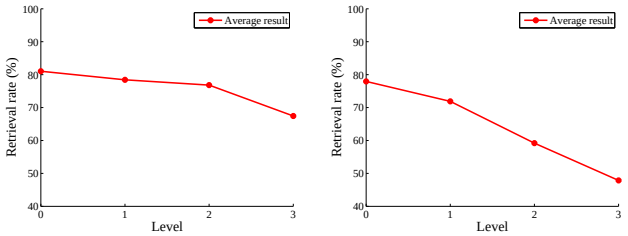
Figure 3 The result of wavelet-based SIFT descriptor (a) and SURF descriptor (b) on each category of the Stanford data.

4.1 The performance analysis of decomposition level l

The decomposition level l will affect the simplicity and the descriptive power of the wavelet-based image descriptor, directly. With small l, the illustrative power of the descriptor will be ensured, but the size of the descriptor will be over larger than necessary, leading to increase the transmission time and lower the user experience. On the contrary, when the value of l is too large, the original descriptor will be decomposed so deep that the new descriptor can not describe the query image at all, which makes the retrieval rate unacceptable. To find the balance between the simplicity and the descriptive power, we set different values of l on the SIFT descriptor and the SURF descriptor, respectively.

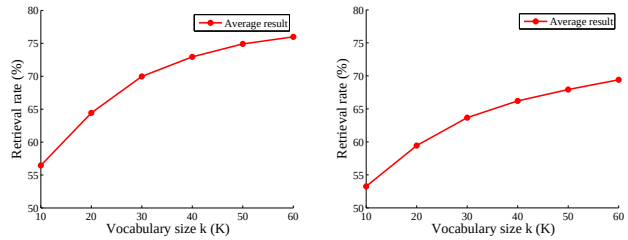
For the SIFT descriptor, the original 128-dimension descriptor is on the level 0. As shown in Fig. 3-a and Fig. 4-a, the retrieval rate is getting lower and lower as l increases. Noticing the result of landmarks category is not shown for its very low value. When $l = 0, 1, 2$, the downtrend of the retrieval rate is linear and not falloff too much. And the average of the retrieval rate is above 70%, which can meet the needs of user experience nearly. When $l = 3$, transmitting 16-dimension wavelet-based SIFT descriptor only, the retrieval rate is significantly reduced and the the average of the retrieval rate drops to 60% in a nose dive. This is to say, setting $l = 2$ is a reasonable way to start the retrieving for the SIFT descriptor, not only reducing the amount of transmitted data and the waiting time of the clients, but also ensuring the retrieval rate in a high level.

Fig. 3-b and Fig. 4-b shows that, for the 64-dimension SURF descriptor, the flex point is $l = 1$. When l is larger than 1, the retrieval rate is about 50% even lower, making it inappropriate for mobile visual search application. The performance of the wavelet-based SURF descriptor is not as good as the wavelet-based SIFT descriptor due to SURF own compressibility of SIFT.



(a) (b)

Figure 4 The average retrieval rate of (a) SIFT descriptor and (b) SURF descriptor for all the categories in the Stanford data.



(a) (b)

Figure 5 The relation between the average retrieval rate and vocabulary size for (a) SIFT descriptor and (b) SURF descriptor.

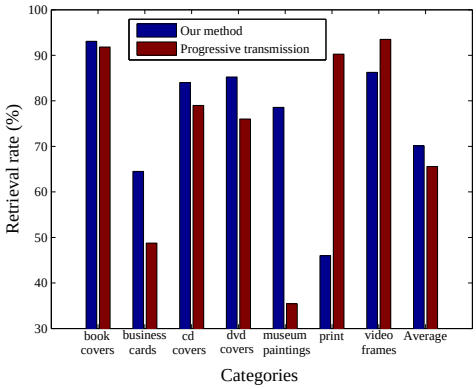
As discussed above, the default values of decomposition level for SIFT and SURF are 2 and 1, respectively.

Table 1 Vocabulary size and Retrieval rate

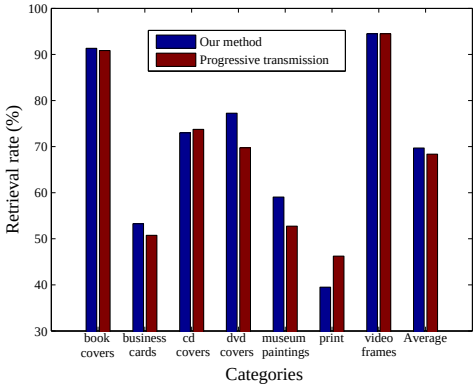
Method	SURF						SIFT					
k	10K	20K	30K	40K	50K	60K	10K	20K	30K	40K	50K	60K
book covers	86.39%	92.08%	93.56%	94.55%	94.80%	95.05%	83.17%	88.61%	93.07%	95.30%	95.79%	95.79%
business cards	40.50%	47.25%	52.75%	53.25%	53.25%	58.25%	45.00%	55.00%	64.50%	69.25%	74.25%	76.75%
cd covers	64.75%	72.75%	77.25%	81.50%	85.75%	85.50%	62.75%	75.75%	84.00%	85.75%	89.75%	92.00%
dvd covers	67.75%	74.25%	79.00%	82.25%	84.00%	85.75%	73.75%	81.00%	85.25%	87.25%	88.50%	88.50%
landmarks	3.99%	5.79%	6.19%	8.38%	8.58%	9.98%	17.96%	22.36%	22.16%	22.75%	21.96%	23.15%
museum paintings	46.70%	54.40%	65.38%	68.68%	72.53%	72.53%	67.31%	74.18%	78.57%	80.49%	83.24%	85.16%
print	28.50%	37.00%	40.50%	44.25%	47.50%	51.25%	34.00%	38.25%	46.00%	52.00%	52.00%	52.00%
video frames	87.50%	92.00%	94.75%	96.75%	97.00%	97.00%	67.75%	80.25%	86.25%	90.75%	93.75%	94.50%
Average	53.26%	59.44%	63.67%	66.20%	67.93%	69.41%	56.46%	64.43%	69.98%	72.94%	74.91%	75.98%

4.2 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 it 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.



(a)



(b)

Figure 6 The comparison between our method and other methods for (a) 32D-SIFT descriptor and (b) 32D-SURF descriptor.

To select a proper k for Stanford data set, we set $k = 10K, 20K, ..., 50K, 60K$ and do the retrieval experiments using the SIFT and SURF descriptor with all the other parameters chosen in the default case. All the results of the eight categories are listed in Table. 1 and the average retrieval rates are shown in Fig. 5. From the result, we can tell the recognition rate grows as the vocabulary size k gets bigger, even when a vocabulary size k is $60K$, the recognition rate still appears to be increasing. This is to say that further gains could be achieved by increasing k.

In the meanwhile, note that the growth rate of the average retrieval rate is getting slower with the growth of the vocabulary size k. And after $k = 30K$, the growth rate of the average retrieval rate has stabilized. Considering the retrieval efficiency, continue to increase k will not bring benefits equivalently.

Thus, $k = 30K$ is a properly chosen for mobile visual search, considering both efficiency and accuracy. So we set $k = 30K$ as the default value of the vocabulary size k in the experiment.

4.3 Retrieval experiments on Stanford data set

In this experiment, we will compare our progressive transmission based on wavelet with other progressive transmission approaches Chandrasekhar et al. (2010) Xia et al. (2012).

In our experiment, we firstly extract the SIFT feature points and SURF feature points on each query image respectively. To keep the same amount of data transmitted via bandwidth constrained wireless link and compare fairly with different kinds of methods, an transmission strategy is designed. Taking the SIFT method for an example, we start with the approximation coefficients of level 2, 32-dimension wavelet-based image descriptors. Thus, the amount of data transmitted is 25% of the original SIFT descriptors. In the contrastive experiment of the progressive transmission approaches Chandrasekhar et al. (2010), we transmit the first quarter 128-dimension SIFT descriptors, which makes its total transmitted data is equal to our method.

As Fig. 6-a shown, our method obtains a higher retrieval rate than the progressive transmission approach Chandrasekhar et al. (2010) in most of the categories, with the same total amount of data transmitted. When the searching result from this level can not meet the needs of the phone users, we transmit the 32-dimension detail coefficients of level 2 to the server side. Then, the 32-dimension approximation coefficients and detail coefficients will be combined to rebuild the 64-dimension approximation

coefficients of level 1, which can be used to search for more matching images.

For the SURF descriptors, we do the same steps as the SIFT ones. From Fig. 6-b, we can tell the retrieval rate of our method is higher than SURF's or equal in most of the categories. Meanwhile, it is interesting to point that our method performs worse than both of the transmission strategy based on SIFT and SURF in the print category, which consists of images captured on newspapers and articles. One reason found by the experiment is that this category tends to achieve higher recognition rate with less descriptor used. Thus, the other two strategies perform better since they only use some descriptors sorted at the front.

5 Conclusions

This paper introduces a way of progressive transmission based on wavelet used in mobile visual search. Due to its progressive transmission mechanism, our method can transmit data efficiently and rapidly via a bandwidth constrained wireless link. Besides this, the capability of restoring the original image ensures user experience under any circumstance. To our knowledge, there is no other proposed strategy that uses both multi-level wavelet decomposition and reconstruction theorem on MVS, which is technically novel. At the same time, to correspond to the multiple levels of wavelet theorem, we built a multi-level bag-of-word system, which is also a shinning point of our strategy. Experiments show that the proposed method outperforms other strategies based on SIFT and SURF descriptors. In our future work, we will focus on combine our method with an application of MVS in a mobile environment to find out the ultimate practicability of our strategy.

Acknowledgment

This work is supported by the National Science Foundation of China under grant nos.of 61173165, 61370199, 61300187, 61370198, New Century Excellent Talents (NCET-10-0095), and the Fundamental Research Funds for the Central Universities under grant nos.of 2013QN029, 2013QN044, DUT12RC(3)89, and DUT13JS04.

References

Leung, CHC and Leung, TLY. (2008) 'Mobile multimedia object management for efficient content search and exchange', *International Journal of Virtual Technology and Multimedia*, Vol. 1, No. 1, pp.61–74

Li, Keqiu. (2011) 'Advanced Topics on Embedded Computing', *International Journal of Embedded Computing*, Vol. 4, No. 1, pp.1–2

Chandrasekhar, Vijay and Chen, David M and Lin, Andy and Takacs, Gabriel and Tsai, Sam S and Cheung, Ngai-Man and Reznik, Yuriy and Grzeszczuk, Radek and Girod, Bernd. 2010 'Comparison of local feature descriptors for mobile visual search', *ICIP*, pp.3885–3888

Li, Keqiu and Shen, Hong and Chin, FYL and Zhang, Weishi (2007). 'Multimedia Object Placement for Transparent Data Replication'. *IEEE Trans. on Parallel and Distributed Systems*, Vol. 18, No. 2, pp. 212–224

Lowe, David G. (1999) 'Object recognition from local scale-invariant features', *ICCV*, pp.1150–1157

Chandrasekhar, Vijay and Takacs, Gabriel and Chen, David M and Tsai, Sam S and Reznik, Yuriy and Grzeszczuk, Radek and Girod, Bernd. (2012) 'Compressed histogram of gradients: A low-bitrate descriptor', *International Journal of Computer Vision*, Vol. 96, No. 3, pp.384–399

Ke, Yan and Sukthankar, Rahul. (2004) 'PCA-SIFT: A more distinctive representation for local image descriptors', *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, Vol. 2, pp.II–506

Bay, Herbert and Ess, Andreas and Tuytelaars, Tinne and Van Gool, Luc. (2008) 'Speeded-up robust features (SURF)', *Computer vision and image understanding*, Vol. 110, No. 3, pp.346–359

Mikolajczyk, Krystian and Schmid, Cordelia. (2005) 'A performance evaluation of local descriptors', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 27, No. 10, pp.1615-1630

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

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

Mulcahy, Colm. (1997) 'Image compression using the Haar wavelet transform', *Spelman Science and Mathematics Journal*, Vol. 1, No. 1, pp.22-31

Sivic, Josef and Zisserman, Andrew. (2003) 'Video Google: A Text Retrieval Approach to Object Matching in Videos', *ICCV*, pp.1470

Vedaldi, Andrea and Fulkerson, Brian. (2010) 'VLFeat: An open and portable library of computer vision algorithms', *Proceedings of the international conference on Multimedia*, pp.1469–1472

Aly, Mohamed and Munich, Mario and Perona, Pietro. (2011) 'Indexing in large scale image collections: Scaling properties and benchmark', *WACV*, pp.418–425