TOP-SIFT: A New Method for SIFT Descriptor Selection

Yujie Liu\textsuperscript{1}  Xiaoming Chen\textsuperscript{1}  Qilu Zhao\textsuperscript{1}
\textsuperscript{1}College of Computer and Communication Engineering  
China University of Petroleum  
Qingdao, China  
liuyujie@upc.edu.cn

Zongmin Li\textsuperscript{1}  Jianping Fan\textsuperscript{2}
\textsuperscript{2}Department of Computer Science  
University of North Carolina  
Charlotte, U.S.A  
jfan@uncc.edu

Abstract—The large amount of SIFT descriptors in an image and the high dimensionality of SIFT descriptor has made problems for large-scale image dataset in terms of speed and scalability. In this paper, we propose a descriptor selection algorithm via dictionary learning and only a small set of features are reserved, which we refer to as TOP-SIFT. We discover the inner relativity between the problem of descriptor selection and dictionary learning for sparse representation, and then turn our problem into dictionary learning. Compared with the earlier methods, our method is neither relying on the dataset nor losing important information, and the experiments have shown that our algorithm can save memory space and increase the retrieval speed efficiently while maintain the recognition performance as well.

Keywords—descriptor selection; dictionary learning; sparse coding

I. INTRODUCTION

SIFT descriptor is the most important local image features recently [1]. Local descriptors are commonly employed in a number of real-world applications such as image classification, wide baseline image matching and registration, and content-based image retrieval.

It has been illustrated that SIFT outperforms other descriptors. The standard local descriptor used by SIFT is constructed by sampling the magnitudes and orientations of the image gradient in the area around the local point, and building smoothed orientation histogram to describe the important information of the area. A 128-element vector is normalized to unit length. SIFT descriptor is a distinctive invariant feature used to robustly depict digital image content. While the large amount and the high dimensionality of SIFT descriptor can bring difficulty for large-scale image dataset in terms of speed and scalability. For example, large-scale object recognition may include comparing of millions of local SIFT descriptors with high dimension, requiring millions of distance computations and millions of storage space.

The solutions to the above problem can be divided into three different approaches. The first is the most widespread, which reduces the dimensionality of the 128-element SIFT descriptors. There are typical works such as PCA-SIFT [2] and SURF [3]. However this approach can lose important information leading to the decrease of accuracy. The second approach is based on the approximate solutions such as LSH [4] and hierarchical structure. This approach can cut down computation time significantly, while it doesn’t work in storage space. Finally, the third approach reduces the number of descriptors which is effective in both computation time and storage space [5][6]. In the present work we address the third approach, which is orthogonal to the other two, and we may combine them in future work.

In this paper we present a method to reduce the number of SIFT descriptors in an image by using dictionary learning for sparse representation. We found that there are many similar or even repeated SIFT points in an image, especially for the complicated or cluttered images. This is the root cause of the large number of descriptors, and many of them are redundant. In this work our aim is to exploit the redundancy in the feature set and select the representative descriptors to be reserved, which we refer to as TOP-SIFT (Fig. 1). In the experiment, we contrast the TOP-SIFTs with the original features in image retrieval to verify the effectiveness of our method.

The rest of this paper is organized as follows: The state-of-the-art techniques for descriptor reduction are
discussed in Section 2. In Section 3, our proposed algorithm for selecting SIFT descriptors is introduced. Our experimental results are shown in Section 4. Finally, the conclusion is summarized in Section 5.

II. RELATED WORK

Most existing algorithms for descriptor reduction can be categories: (1) reserve object descriptors; (2) keep the top N descriptors by using some criterion; (3) sparse sampling images to reduce descriptors.

For the first category algorithm, the generalized strategy is to identify the commonalities in appearance amongst a group of images and keep the descriptors on the common regions. However, these algorithms always depend on the datasets, and some are supervised, which limit their applications and promotion to a large extent. Knopp et al. [9] have developed a method to remove the confusing features in place recognition by computing the confusion score in a sliding window fashion. However, it assumed that images are already geolocated, which is limited in other domains. Turcot et al. [5] proposed that the points which are likely to be found in more than one image of the same object or location are useful features. The approach has removed lots of descriptors and kept the performance, yet it relies on the given images and is not suitable for unrelated image sets.

For the second category algorithm, the common strategy is to analyze the property of the feature points based on some criteria and select small subsets. Foo et al. [6] proposed a descriptor selection method based on the rank of contrast values, and save top N features, while it is likely to lose important information. In [10] Brown et al. presented an adaptive non-maxima suppression method to select a fixed number of interest points with a good spatial distribution in the image. This approach is orthogonal to many methods and can be combined in next work.

For the third category algorithm, sparse sampling images to reduce the descriptors indirectly. It is mainly used in 3D reconstruction. However it is proved that we will get better results in 3D reconstruction through dense-sampling of the images [11]. Moreover, in many situations, the obtained set of descriptors could still be very large, such as reconstructing 3D scenes using thousands of images [8].

III. OUR PROPOSED ALGORITHM

In this section, we define the criterion to value the SIFT descriptors and reformulate the problem of descriptor selection via dictionary learning under sparsity constraint. We firstly explain the main idea of our problem reformulation, and then introduce the dictionary learning method to address the problem.

A. Problem Reformulation

In the traditional method of SIFT descriptor detection, it produces many descriptors, especially for the cluttered or complicated images. However, we discover that there exist many descriptors being similar or even the same. Therefore, the large number of descriptors is redundant in fact. Our aim is to remove the redundant descriptors and find the representative ones which we refer to as TOP-SIFTs. The

\[
\text{TOP-SIFTs should have the ability to represent all the other ones. So this problem can be regarded as reconstruction problem, which all the descriptors can be reconstructed by the TOP-SIFTs. To obtain the biggest reconstruction ability we should minimize the overall reconstruction error given as:}
\]

\[
\min_D ||X - f(D)||^2
\]

where X is the original SIFT set and D is the set of TOP-SIFTs. The definition of function f(.) determines how each descriptor in the original SIFT set can be reconstructed by the TOP-SIFTs. In this paper, we define the function f(.) to make D sparsely reconstruct each descriptor in the original set X. Therefore, we reformulate the task of descriptor selection as the problem of dictionary learning for sparse representation. The two problems are linked together by their inner relativity: both of them attempt to select a subset that can reconstruct large number of items in the original set.

Finally, considering the representation ability of one SIFT descriptor is limited. We add sparsity constraint to the objective function so that limited amount of coefficients will be effective in the reconstruction. Therefore, only the bases whose coefficients are non-zero will take effect in the reconstruction. The final objective function is as follows with sparsity constraint.

\[
\min_{D, \alpha} \sum_i ||x_i - D \alpha_i||^2 + \lambda \sum_i ||\alpha_i||_0 \quad \text{s.t. } \alpha_i \geq 0
\]

where \(x_i\) is the original 128-element descriptor and \(\alpha_i\) is the sparse coefficient. For each image we will learn a dictionary D whose bases are the representative descriptors. 

B. Dictionary Learning

Although descriptor selection and dictionary learning for sparse representation have inner relativity, we must clarify that our method has its distinctiveness. On the one hand, during the process of dictionary learning, the bases selection range is fixed, namely, the dictionary is constructed by real samples. On the other hand, the coefficients are nonnegative, the sparse coefficients either be zero or positive. Because it is impossible for negative coefficients to reconstruct the original SIFT feature set. We are inspired by the method [12] for dictionary learning, and adopt simulated annealing algorithm to solve our optimization problem.

Simulated annealing algorithm is stochastic searching optimization algorithm based on the Metropolis criterion, which can get the approximate global optimal solution. The cooling schedule we accepted is defined as below:

\[
T_k = \frac{T_0}{\log (k+1)}
\]

where \(T_0\) and \(T_k\) are the initial and current temperature respectively, \(k\) is the iterative time. The temperature decrease faster during the initial steps and slow down later, which make the acceptance condition relax in the beginning and we can search the optimal solution sufficiently, then turn to rigorous at the later stage for avoiding get poor solution ultimately.
Traditional acceptance probability is set as $P = \exp\left(-\frac{R(D_k)-R(D_{\min})}{\alpha R(D_{\min}) T_k}\right)$, we reform it to fit our problem better, which is defined as below:

$$P = \exp\left(-\frac{R(D_k)-R(D_{\min})}{\alpha R(D_{\min}) T_k}\right) \quad (4)$$

The SIFT descriptors vary greatly with different images, we add the current optimum to denominator in order to keep the same criterion for every image to search the optimal solution. For different original descriptor sets, the acceptance probability defined in (4) can keep a near-identical pace to determine if the new solution is accepted. The optimum accepted principle is defined as follows:

$$D_{\min} = \begin{cases} 
D_k & R(D_k) < R(D_{\min}) \\
D_k & R(D_k) \geq R(D_{\min}) \text{ and } P > U \\
D_{\min} & R(D_k) \geq R(D_{\min}) \text{ and } P \leq U 
\end{cases} \quad (5)$$

where $R(.)$ is the reconstruct function defined in (2) and $D_{\min}$ is the optimal solution, $U$ is a random number between 0 and 1. We accept the random mechanism of simulated annealing algorithm to get the approximate global optimum.

Our method of basis updating refers to [12], which each updated base is a real SIFT sample. The initialization of the dictionary is selecting a fixed number of SIFT descriptors randomly as the bases. For each base we compute its similarity with the other descriptors in the original descriptor set, and rank the value in descending order. Then we search the new base in the neighborhood of the old base, which the search range is limited by $\exp\left(\frac{T_k-T_0}{T_0}\right)[X]$. We link the search range and temperature together, which is beneficial to find the candidate sufficiently in the earlier steps and keep the optimum later. We must make sure that there are no identical bases in the new dictionary. After all the bases are updated, we calculate sparse coefficients as the method of [13] and compute the new reconstruction error as (2). Whether the new dictionary will be received is determined by the accepted principle defined as (5), and the dictionary accepted will be used as the input for next iteration. Next, the dictionary and the coefficients are updated iteratively until the temperature reaches the predefined low value or the reconstruction error is not being updated for maximum consecutive rejection times. An overview of our method is presented in Algorithm 1.

### Algorithm 1: proposed dictionary learning for SIFT descriptor selection

**Input:** Original SIFT features set $X^{128 \times m}$.

**Output:** Optimal dictionary $D^{128 \times n}$.

**Initialization:** The initial dictionary is constructed by random selection of $n$ bases from $X$.

**Dictionary Learning:**

while $T_k > T_{\text{stop}}$ and $R_{ej} < \text{MaxConseRef}$ do

1. $D_k = \text{Update}(D_{k-1})$;
2. $A_k = \text{Sparse}(D_{k-1}, X)$;
3. $R_k = \text{Compute}(D_k, A_k, X)$;
4. if accept($R_k, D_k$) then
   1. $R_{\text{opti}} = R_k$;
   2. $D_{\text{opti}} = D_k$;
   3. $R_{ej} = 0$;
5. else
   1. $R_{ej} = R_{ej} + 1$;
   2. $k = k+1$;
   3. $T_{k+1} = \text{Update}(T_k)$;

end if

end while

The SIFT feature extraction algorithm we employed is [7]. We use the traditional Bag-of-words (BOW) method to quantize the descriptors into “visual words” and adopt tf-idf weighting scheme. The vocabulary size is 20K trained by ourselves. The baseline algorithm is the above retrieval model with all SIFT descriptors, and then we verify the proposed descriptor selection method with the same retrieval structure. We adopt mean average precision (MAP) as the evaluation criterion, which is widely used in image retrieval.

### B. Experimental Results

To verify our proposed method we select different proportional descriptors and apply them into image retrieval. The results are summarized in Fig. 2 and Fig. 3. For some images, the amount of the descriptors is small and all of them are reserved.

Fig. 2 shows the average query time and memory space with different proportional SIFT descriptors. As we can see by the curves, the memory space decreases sharply with the reduction of descriptors, so it is with the query time. Fig. 3 shows the overall trend of MAP with different proportional descriptors. From the curve we can observe that the retrieval performance keeps stable until the descriptors reduced to about 30%. To store the total descriptors would roughly require 1.3GB memory space, however it would only require 0.75GB for 50% descriptors. It is also effective in average query time and nearly half of the original cost is saved. What’s more, the MAP just decreases slightly, not more than 1%. For TOP-SIFTs with 40% and 30%, the MAP decreases a little within the acceptable range, while we get improvement in both the time and memory space.

238
significantly. In conclusion, though SIFT descriptors are dramatically pruned, the results show that the effectiveness is almost negligible lost compared with the original set. This fully illustrates that the original feature set is redundant, there exit lots of useless descriptors and waste large memory space. For 20% descriptors, the MAP decreases a little more, this is because the removed descriptors are so much that the reserved features can’t represent the image sufficiently. So the descriptor reduction is not arbitrary, the selection of TOP-SIFTs should keep appropriate proportions. Furthermore, the overall precision of our experiment is lower than the state-of-art results, which is closely related to the retrieval approaches. In this paper we just want to verify the traditional SIFT descriptors are redundant and our method for descriptor selection is effective. In the future work, we will combine the state-of-art image retrieval techniques with our descriptor selection method, then we can not only remove the descriptors redundancy but also improve the overall recognition performance.

**Figure 2.** Average query time (secs) and memory space (GB) using different proportional SIFT descriptors on Holidays dataset.

**Figure 3.** Mean Average Precision of retrieval answers using different proportional SIFT descriptors on Holidays dataset.

V. CONCLUSION

In this paper, we present a SIFT descriptor selection method via dictionary learning. Earlier methods proposed for descriptor selection either rely on the dataset, which is always inefficient for a single image, or are easy to lose important information. We discover the inner relativity between the problem of descriptor selection and dictionary learning for sparse representation, and then turn our problem into dictionary learning. In the experiment our proposed algorithm can save memory space, decrease consuming time efficiently, and maintain the recognition performance as well. In the aspect of scalability, our method can be employed into all the areas in which SIFT features are involved. In the next work, we will apply our method to more areas, not just image retrieval, but also 3D reconstruction, image clustering, etc. Then it will be more meaningful.

ACKNOWLEDGMENT

This work is partly supported by National Natural Science Foundation of China(61379106), the Scientific Research Foundation for the Excellent Middle-Aged and Young Scientists of Shandong Province of China (Grant No.BS2010DX037), the Shandong Provincial Natural Science Foundation (Grant No.ZR2009GL014, Grant No. ZR2013FM036), the Open Project Program of the State Key Lab of CAD&CG (Grant No. A1315), Zhejiang University, the Fundamental Research Funds for the Central Universities(Grant No.14CX06081A,14CX06010A, 14CX06012A). Innovation Project of China University of Petroleum (Grant No. YCXX14066).

REFERENCES