A Feature Compression Scheme for Large Scale Image Retrieval Systems

Umair Mateen Khan
Computer Science Dept.
Otago University
Dunedin, New Zealand
umairkhan@cs.otago.ac.nz

Brendan McCane
Computer Science Dept.
Otago University
Dunedin, New Zealand
mccane@cs.otago.ac.nz

Andrew Trotman
Computer Science Dept.
Otago University
Dunedin, New Zealand
andrew@cs.otago.ac.nz

ABSTRACT

Many image retrieval and object recognition systems rely on high-dimensional feature representation schemes such as SIFT. Because of this high dimensionality these features suffer from the curse of dimensionality and high memory needs. In this paper we evaluate an approach that reduces the size of a SIFT descriptor from 128 bytes to 128 bits. We test its performance in an image retrieval application and its robustness in the presence of various image transformations. We also introduce and evaluate a simpler approach that requires no training but requires 512 bits per descriptor.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Compression, Feature Measurement, Applications; I.5 [**Pattern Recognition**]: Clustering, Applications; G.1 [**Numerical Analysis**]: General

General Terms

Measurement, Performance, Experiments

Keywords

Feature compression, Large scale image retrieval, Curse of dimensionality.

1. INTRODUCTION

With the advent of robust low level features, image retrieval and object recognition has improved considerably over the last several years. Feature extraction algorithms like [10, 1, 11, 16] are used to extract keypoints from an image and then represent the information in the form of high dimensional feature vectors. Because of this high dimensionality these features suffer from the curse of dimensionality and have high memory requirements. The size of the descriptor is critical in memory limited applications such as those involving mobile or embedded devices. One possible solution is to generate fewer features. Unfortunately, fewer features results in significantly worse performance for image retrieval applications [7]. The alternative is to keep many features but reduce the memory footprint of each feature. In this paper we evaluate the approach of [13] that reduces the size of the descriptor down to just 1 bit per dimension i.e. 128 bits for a standard SIFT [10] feature vector and uses the Hamming distance for feature comparison. The value in each dimension is thresholded to a binary value based on a data dependent threshold. Effectiveness of the Hamming distance for object recognition scenarios is evident in previous work [13]. We have evaluated the 1-bit SIFT descriptor in an image retrieval application and have also tested its robustness to different transformations e.g. illumination changes, rotation and blurring etc. We have also tested a method that simply chooses the most significant bits of each dimension to represent a feature. We also compare these two approaches against the original SIFT descriptor. The research questions that we answer in this paper are:

- Is a 128-bit SIFT descriptor appropriate for image retrieval applications?
- Can a simpler approach which does not require training perform comparably?
- How do these descriptors perform under different image transformations?

Related Work

The concept of feature size reduction, dimensionality reduction and curse of dimensionality are not new and are interlinked with each other. In PCA-SIFT [16] Principal Component Analysis (PCA) was used to reduce the size and dimensions of the of the SIFT descriptor. PCA was applied on the normalized gradient patch across each keypoint to reduce the descriptor to just 36 dimensions and the descriptor was capable of very high performance. However using PCA requires off-line training to estimate the covariance matrix used for PCA projection. BRIEF [3] uses a very short binary string descriptor based on Naive Bayes comparison of image patches using 256 or 128 bits. The proposed descriptor was fast and competitive to SURF and U-SURF[1]. In [15] SIFT descriptors were reduced to just 36 dimension by applying kernel projection on the orientation gradient patches instead of using smoothed weighted histogram.The descriptor was short but very effective and was tolerant to geometric distortions. The approach was named KPB-SIFT and does not require a training stage.

IVCNZ '12, November 26 - 28 2012, Dunedin, New Zealand

In [9] the number of SIFT features and the feature size were reduced by ignoring rotational invariance - an appropriate choice for indoor environments. In [4] a technique based on transform coding was presented. They showed that SIFT and SURF descriptors can be reduced to below 2-bits per dimension providing a compression rate of 16 times relative to the conventional floating point representations. Features are encoded by first applying PCA and then scalar quantizing each dimension using arithmetic coding. An inverse process is applied during decoding time. The approach produces good performance using 57-bits per descriptor and only resulted in negligible image matching error. In [13] and [14] a binary descriptor with just 1 bit per dimension was introduced using the median value as the threshold. In [6] a descriptor compression approach that does not need decompression during matching time was introduced. The feature size was reduced by an order of magnitude and yet achieve a detection rate of 95%. They converted SIFT, SURF and GLOH into a canonical form that showed better results than the original descriptor. Brown et al [2] introduced a descriptor learning technique that used both linear and non-linear dimensionality reduction and also used discriminant learning techniques (LDA) and optimisation methods to find the optimal parameters. The number of bits required was further reduced to 2 per dimension and still a good error rate was maintained. They also suggested the need for a variable number of bits for each dimension as the variance on each dimension can differ substantially across the descriptor.

In all previous work the size and dimensionality reduction was achieved by either following a complex preprocessing step or a training stage used to learn different parameters that are then used to produce a reduced descriptor. The purpose of the work presented here is quite different, although the technique used is similar. Our goal is to show how well a simple feature reduction approach can perform without the need for any training. Our second goal is to test an already proposed approach in an image retrieval scenario. We then compare these approaches with the normal SIFT descriptor.

2. COMPRESSION TECHNIQUES

We have compared four different compression techniques and the original SIFT descriptor:

SIFT-6 Use the 6 most significant bits per dimension.

SIFT-4 Use the 4 most significant bits per dimension.

SIFT-2 Use the 2 most significant bits per dimension.

SIFT-1 Use the method of [13] to compress to 1 bit per dimension.

SIFT-2, SIFT-4, and SIFT-6 simply discard the least significant bits and hence do not require training. However, the method does require packing and unpacking bits at encoding and decoding time, as the Euclidean distance is still needed for feature matching. For these descriptors we use an empirically determined matching threshold of 210. In contrast, SIFT-1 requires estimation of the median for each dimension from a training collection. However, after encoding is done, the Hamming distance can be used for feature matching. Figure 1 gives a pictorial representation of each scheme. An experimentally determined Hamming threshold of 25 was used to determine if two features matched.

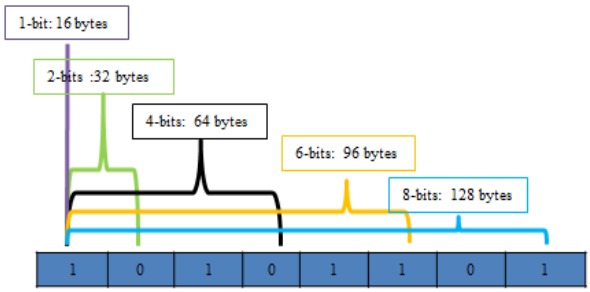


Figure 1: The suggested feature compression approach is illustrated. The total descriptor size is also displayed in bytes. Here x-bits means the number of bits kept from each dimension of the descriptor.

3. RESULTS

Three different sets of experiments are performed. Section 3.1, describes the benchmark datasets used in these experiments. Firstly the compression methods are tested for robustness to several image transformations and this is explained in Section 3.2. Then in Section 3.3, the methods are tested using real image retrieval datasets. Finally, in Section 3.4 the methods are tested using a real image retrieval dataset whilst using 1 image per scene for training. Notice that in all these experiments a retrieved image is considered a match if a correct image is the top ranked image i.e. image having maximum score. The matching score is the number of terms (features) matched with the query image. Also, *Accuracy* is the performance measure used for all experiments and is described in the Equation 1.

$$Accuracy = \frac{Total\_Correct\_Matches}{Total\_Query\_Images} \quad (1)$$

3.1 Benchmark Datasets

UK Bench Dataset [12]

This dataset contains 10,196 images and there are four images per scene in the dataset possibly with different transformations i.e scale, rotation, illumination changes and view point changes. For our experiments we used a maximum of first 1000 scenes i.e. 4000 images. The first three images of every scene are kept for training purposes and the fourth image is selected as query image as shown in Figure 10(a).

INRIA Holiday Dataset

This Dataset [5] contains 1491 images. Out of these, 991 images were selected for training and 500 images were used for testing. Some query images have just one training image while others have multiple training images in the dataset. Some images of this dataset are shown in Figure 10(b).

Otago University Dataset

This dataset contains 2000 Images of both indoor and outdoor scenes. The dataset was prepared in the same manner as UK Bench dataset[12] i.e. there are 500 scenes in total with four images per scene. Here also, the first three images of every scene are kept for training and the fourth image is selected as query image. The images contain different transformations like rotation, translation, view point

change, scaling and illumination changes and is much more challenging compared to UK Bench Dataset. Some images of this dataset are shown in Figure 10(c). Notice the amount of viewpoint change in these images.

3.2 Performance against Transformations

For testing the performance of these three approaches against different transformations we used the same set of experiments as used in [8]. The experiments are carefully designed to test the matching performance in different scenarios like rotation, blurring, illumination changes, noise, viewpoint and scale changes. In all of these experiments the first 500 scenes from the UK Bench dataset [12] are used as the image collection. For testing, each of the collection images are transformed and the image with the most matching features is selected as the matched image. There is only one correct match in the collection for each test image. We started by testing the matching performance in the presence of rotation in the images. The trained images were rotated by 40, 135, 215, 250 and 300 degrees. Notice that these rotated versions were generated by rotating the training images in Matlab. From Figure 2 it can be seen that rotating does not cause any degradation in performance for any of the techniques.

The next experiment used Gaussian blurring at three different levels of smoothing: $\sigma = 5$, 10 and 20. The results are shown in Figure 3 and show that all methods are robust to moderate levels of blurring and do not degrade until the blurring becomes extreme. Even in this case the accuracy of SIFT and SIFT-6 remained higher then 90%.

The results of illumination changes are shown in Figures 4 and 5. In this case, a constant illumination value was either added or subtracted to each trained image. Again, all methods perform well even up to quite large changes in illumination.

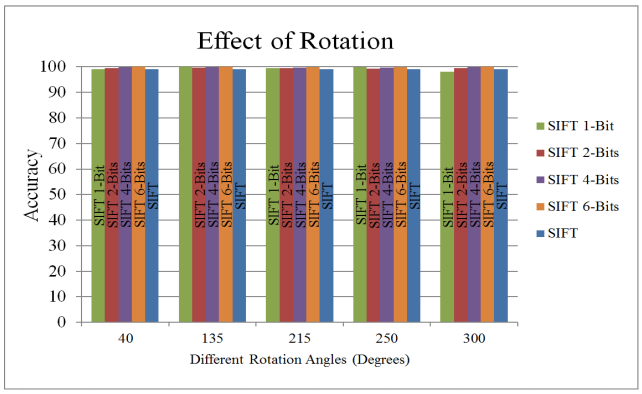


Figure 2: Matching performance in the presence of five different rotations angles applied to the train images.

Figure 6 shows results for various types and level of image noise. Salt and pepper noise produced the biggest reduction in performance with SIFT-6 being the most robust method. Finally, Figure 7 shows results for scale and viewpoint changes. Notice that in each of these cases the images with different scales and viewpoints were chosen from the UK Bench dataset manually by looking at the images from the first 1000 scenes as mentioned in [8]. Figure 7 depicts the results of both scenarios in one chart. Notice the significant amount of decrease in the performance of all methods

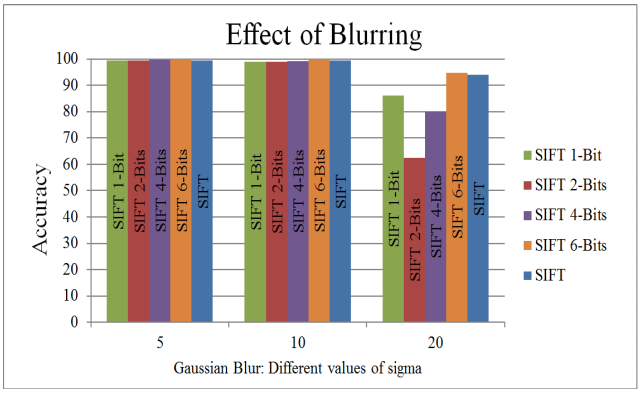


Figure 3: Matching performance against blurred images of three different sigmas (represented by x-axis) in Gaussian blur.

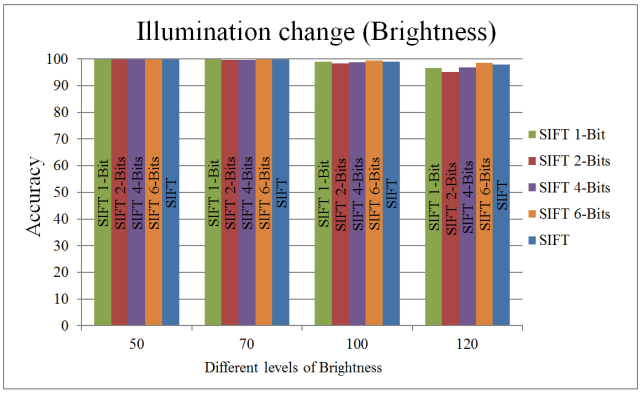


Figure 4: Matching performance of different reduction approaches compared with SIFT descriptor in the presence of brightness or addition of light.

specially in the presence of too much viewpoint changes.

3.3 Image Retrieval Scenario

In this section we test the methods against three real datasets in an image retrieval scenario (see Figure 10). The statistics about these datasets are described in the table 1 and the results are shown in Figure 8. From the results we can see that all approaches performed well while SIFT-1 method performed slightly better than other methods. Notice the drop in the accuracies of all methods in the case of Otago University dataset that reflects how challenging this dataset is.

3.4 One Training Image

Table 1: Three Benchmark Datasets used for checking the image retrieval results.

Dataset	Train	Test	Total Images
UK Bench	3000	1000	4000
INRIA Holiday	1491	500	1991
Otago University	1500	500	2000

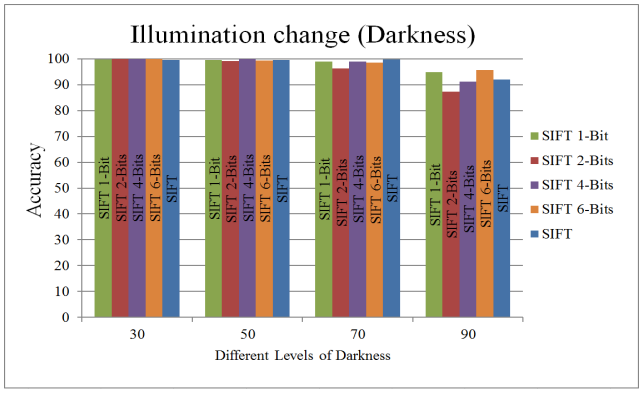


Figure 5: Matching performance of different reduction approaches compared with SIFT descriptor in the presence of darkness or reduction of light.

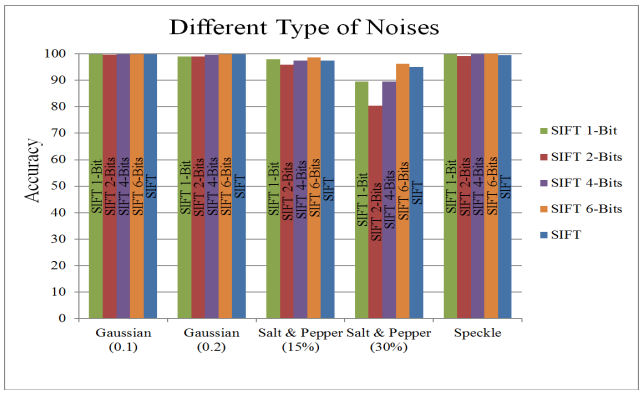


Figure 6: Matching performance of different reduction approaches compared with SIFT descriptor in the presence of different noise.

The final experiment was designed to check the matching accuracy and descriptors distinctiveness in a scenario where only a single training image is available. The experiment was conducted on the first 1000 scenes of the UK Bench Dataset. The first image of every scene (i.e. there are 4 images per scene) was taken as the training image. Image number 4 of every scene was chosen as the testing image. The matching results are depicted in Figure. 9. We can see that SIFT-1 and SIFT-4 actually performed the best but a drop in accuracy of at least 8% was observed for the remaining approaches.

4. CONCLUSIONS

In this paper we investigated the effect of various image transformations on compressed descriptors in an image retrieval application. We also evaluated a simple feature reduction approach that does not require training.

We have found that the 128-bit SIFT is a competitive approach with such a small memory footprint. The key to its strength lies in the way the threshold is chosen to cluster the real values to binary. Choosing the median is key as it was found in [13] that descriptor values are not symmetrically distributed and that many of the values occur in the least significant bits. The approach is fast because the Hamming distance can be used for feature comparison. The only disadvantage is that the method requires a training collection.

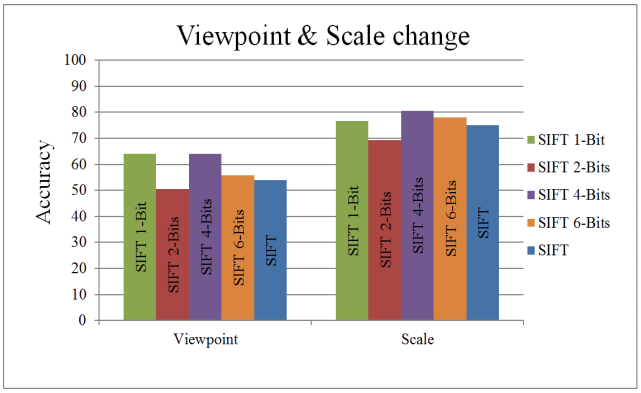


Figure 7: Matching performance of different reduction approaches in two different scenarios i.e. viewpoint and scale change.

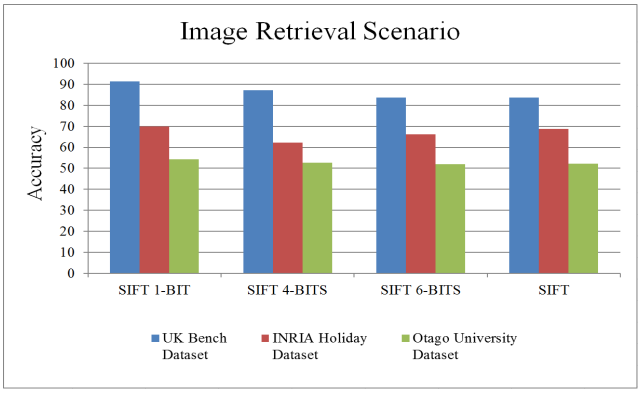


Figure 8: Matching performance of different reduction approaches in two different scenarios i.e. viewpoint and scale change.

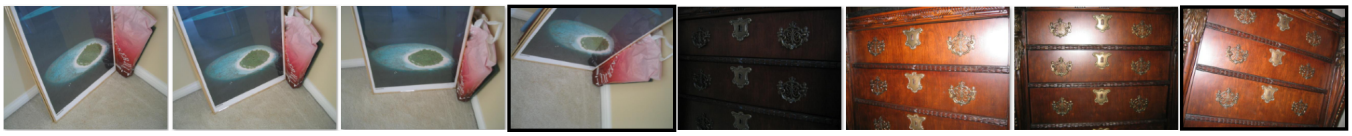
In the case of our approach the results are equally competitive and the method does not need any training at all. The approach however is slower because the bits require packing and unpacking and the need for using Euclidean distance as a distance measure. It is worth noting that in the case of SIFT-4 where we reject the lower 4-bits, that according to authors of SIFT-1 actually contain 50% of the values in the the descriptor dimensions the methods still performed considerably well. All of these methods were found to be robust under different image transformations.

5. REFERENCES

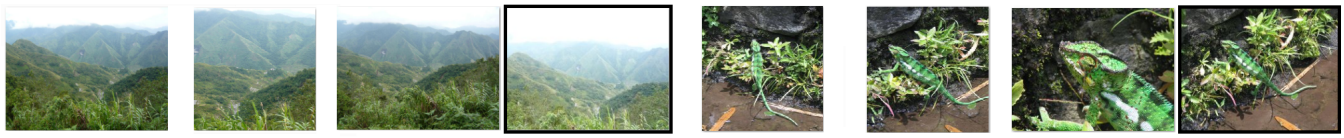
[1] H. Bay, T. Tuytelaars, and L. V. Gool. SURF: Speeded Up Robust Features. In *ECCV*, pages 404–417, 2006.

[2] M. Brown, G. Hua, and S. A. J. Winder. Discriminative learning of local image descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):43–57, 2011.

[3] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. BRIEF: Binary Robust Independent Elementary Features. In *Proceedings of the 11th European conference on Computer vision: Part IV*, ECCV'10, pages 778–792, Berlin, Heidelberg, 2010. Springer-Verlag.



(a) UK Bench [12] dataset



(b) INRIA Holiday [5] dataset

(c) Otago University dataset

Figure 10: Images from our evaluation datasets are shown. The black bordered image is used as query image.

Figure 9: Matching performance of all different approaches in a scenario when there was only one train image in the database against each query image.

[4] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, J. Singh, and B. Girod. Transform coding of image feature descriptors.In *Proc. SPIE 7257, Visual Communications and Image Processing 2009, 725710 (January 19, 2009)*, doi:10.1117/12.805982.

[5] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag.

[6] M. Johnson. Generalized descriptor compression for storage and matching. In *Proceedings of the British Machine Vision Conference*, pages 23.1–23.11. BMVA Press, 2010. doi:10.5244/C.24.23.

[7] N. Khan, B. McCane, and S. Mills. Feature set reduction for image matching in large scale environments. In *IVCNZ*, 2012.

[8] N. Y. Khan, B. McCane, and G. Wyvill. SIFT and SURF performance evaluation against various image deformations on benchmark dataset. In *DICTA*, pages 501–506, 2011.

[9] L. Ledwich and S. Williams. Reduced SIFT features for image retrieval and indoor localisation. In *Australian Conference on Robotics and Automation*, 2004.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

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

[12] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, pages 2161–2168, Washington, DC, USA, 2006. IEEE Computer Society.

[13] M. Stommel and O. Herzog. Binarising SIFT descriptors to reduce the curse of dimensionality in histogram-based object recognition. *International Journal of Signal Processing*, 3(1):25–36, 2010.

[14] M. Stommel, M. Langer, O. Herzog, and K.-D. Kuhnert. A fast, robust and low bit-rate representation for SIFT and SURF features. In *IEEE International Symposium on Safety,Security and Rescue Robotics, Kyoto, Japan,2011*, 2011.

[15] G. Zhao, L. Chen, G. Chen, and J. Yuan. KPB-SIFT: a compact local feature descriptor. In *ACM Multimedia*, pages 1175–1178, 2010.

[16] S. Zickler and A. A. Efros. Detection of multiple deformable objects using PCA-SIFT. In *AAAI*, pages 1127–, 2007.