PQ-WGLOH: A BIT-RATE SCALABLE LOCAL FEATURE DESCRIPTOR

Chunyu Wang, Ling-Yu Duan, Yizhou Wang, Wen Gao

The Institute of Digital Media, School of EE&CS, Peking University, Beijing, 100871, China
{wangchunyu, lingyu, yizhou.wang, wgao}@pku.edu.cn

ABSTRACT

In this paper, we propose a compact yet discriminative local descriptor which tackles the wireless query transmission latency in mobile visual search. The descriptor captures gradient statistics of canonical patches over a log-polar location grid whose parameters are optimized using training samples. We quantize the resulting descriptor using product quantization. The descriptor achieves about 95% bits reduction compared with 128-Byte SIFT and allows adaptation of descriptor lengths to support user required performance. Moreover, accurate matching of descriptors with low complexity is allowed within several table lookup operations. We perform a comprehensive comparison with SIFT, GLOH and CHoG in the context of image retrieval, image matching and object localization. We achieve competing matching and retrieval performance with SIFT, GLOH with much fewer bits. In particular, the descriptor outperforms CHoG at the same bits on eight data sets contributed to MPEG Compact Descriptor for Visual Search(CDVS) Standardization.

Index Terms— Local descriptor, compact, product quantization

1. INTRODUCTION

With the popularization of mobile embedded camera, there is a great potential for mobile visual search. In a typical scenario, the query image descriptors transmission is over a bandwidth constrained wireless network. Thus visual descriptor is expected to be compact, discriminative and meanwhile efficient for matching to tackle the server response latency. To improve descriptors' discriminative power, state of the arts [1, 2] divide canonical patches into several spatial parts and calculate a sub-descriptor for each part. However, the resulting descriptors have high dimensions which undoubtedly increase the query transmission latency.

Robust local feature descriptors have been extensively studied in recent years. Scale Invariant Feature Transform(SIFT)[1] is probably the most widely used one. Bay et al present Speeded Up Robust Features(SURF)[2] which allows fast computation of box type convolution filters with integral images. Gradient Location Orientation Histogram(GLOH)[3], extending SIFT with log-polar location grid is experimentally shown to be more robust than SIFT. Brown et al[4] propose a general framework for optimizing parameters with descriptor computation using direction set method.

More recent emphasis has been put on low bit rate descriptors which can be classified into three groups. The first group applies dimension reduction. Ke[5] performs PCA on gradient patches and only retains dimensions with large variance. Brown et al [4]use Linear discriminant analysis(LDA) to map descriptors to a lower dimensional space where match descriptors are drawn close while non-match descriptors are kept apart.

Representing an image as a binary code is another popular practice in scalable image search. Torralba[6] uses machine learning techniques to convert the GIST descriptor to a compact binary code to enable large scale image search. Strecha[7] proposes to multiply descriptors by a projection matrix, being subtracted from a threshold vector, and finally retain only the sign of the results. The projection matrix and threshold vector are obtained by minimizing the in-class variance and maximizing the covariance between classes as well.

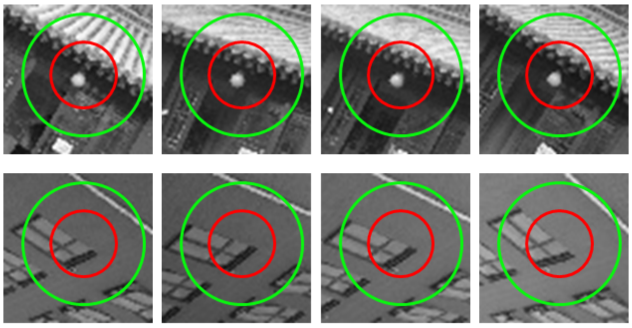


Fig. 2: Each row shows 4 projected patches from a 3D cloud point. Pixels in inner regions vary slightly while outer regions vary largely.

The third class of techniques rely on quantization. In[8], SIFT descriptor is quantized using only 4 bits per dimension, thus reducing the descriptor size to 512 bits. CHoG[9] applies tree-coding to the lossy compression of probability distributions of SIFT-like descriptors. Patch level experiment shows CHoG can largely reduce bit rates but not deteriorate its discriminative power.

Towards low bit rate mobile visual search, we propose a compact yet discriminative local descriptor PQ-WGLOH. It captures gradient statistics from the canonical patches over a log-polar location grid. First, we learn an optimal weight for each dimension of the descriptor and obtain a new descriptor named as WGLOH(Weighted GLOH). Second, with product quantization, we divide WGLOH into several segments and quantize each segment independently. Different from CHoG, the reproduction values of each quantizer are learned in a data driven manner. We refer to our product quantized descriptor as PQ-WGLOH(Product Quantized WGLOH).

Section 2 presents the proposed descriptor. In section 3, we discuss the experimental results. Section 4 concludes the work.

2. THE COMPACT DESCRIPTOR: PQ-WGLOH

We first construct a circular region r with radius R around each interest point(Fig.1.a). Then the region r is partitioned into k subregions r_i by a log-polar grid. We learn three optimal radiuses of the log-polar grid, R_1, R_2 and R_3 in a discriminative way(Fig.1.b). For each sub-region r_i, magnitudes and orientations of image gradients are

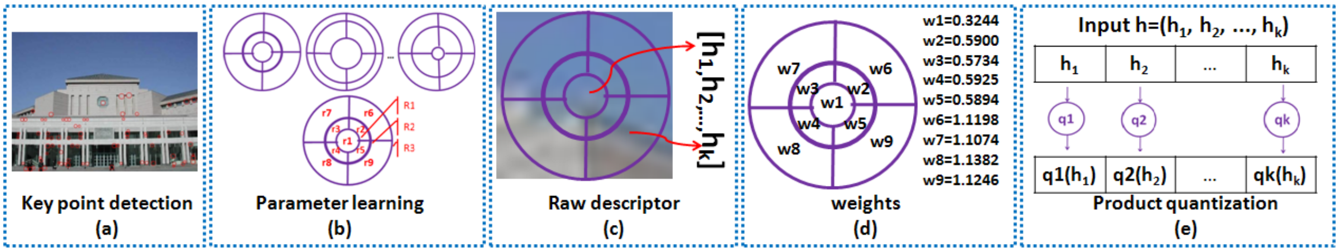


Fig. 1: Flow chart of the proposed descriptor PQ-WGLOH.

accumulated to an orientation histogram h_i with m bins. The local descriptor h of the region r is finally composed of k m-dimensional sub-region histograms: $h = (h_1, ..., h_k)$(Fig.1.c). We refer to each sub-region histogram h_i as a *location bin* of h. We assign a weight to each *location bin*(Fig.1.d) and quantize the weighted *location bin* with product quantization(Fig.1.e).

2.1. Radius R_1, R_2 and R_3 learning

Brown[4] did patch level experiments to learn proper parameters of descriptors. However, the descriptor that works well at the patch level would not guarantee best performance at image level. Hence we directly learn R_1, R_2 and R_3 using image matching experiment.

We first extract local descriptors from each pair of images and match them using the ratio test criterion[1]. With the number of matched descriptors, we form two histograms, one for match image pairs and the other for non-match image pairs. Integrating the two histograms, we obtain a Receiver Operating Characteristic(ROC) curve which plots correct match fraction against incorrect match fraction. The area under the ROC curve is used to measure descriptor performance. To maximize the ROC area, we jointly search optimal R_1, R_2, R_3 by multidimensional direction set method. Experiment on PKU data set[10] shows that $R_1 = 20\% * R, R_2 = 40\% * R, R_3 = 100\% * R$ yields best result. We fix R_1, R_2, R_3 for remaining experiments in this paper.

2.2. Location bins weighting

We learn good choices of weights for each *location bin* using a training set consisting of patches from a multi-image 3D reconstruction where accurate ground-truth matches are determined[4]. Each recovered 3D point P_i is projected back to images in which they were matched to produce accurate virtual interest points. We extract a patch s_{ij} of 64*64 pixels around the virtual interest point in image I_j. Sample patches are shown in Fig.2.

We discover the sub-regions of matched patches have different degrees of variance. Intuitively, the descriptor will be more robust if the variance of matched descriptors is minimized. It encourages us to weight sub-regions by a decreasing function of corresponding variance. We first calculate variance v_{ij} of the j_{th} *location bin* for descriptors of the virtual interest points projected from i_{th} 3D point, which is the sum of variances of all dimensions of the j_{th} *location bin*. Variance v_j of the j_{th} *location bin* for all patches is set to the average of v_{ij}, $v_j = \frac{1}{N} * \sum_{i=1}^{N} v_{ij}$. N is the number of 3D points.

We conducted simulation experiments on PKU data set where $40,000$ 3D points are generated. The result is shown in Fig.1.d which tells us location bins at the same radius almost have the same amount of variance. Innermost region has the least variance while the 4 middle regions have the largest variance. We propose to put the weight of the j_{th} location bin as $w_k = exp(-v_j)$.

2.3. Descriptor quantization

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding where all components of a vector are quantized simultaneously. A disadvantage of VQ is that its encoding complexity increases dramatically with the vector dimension[11].

Product quantization addresses the problem by dividing an input vector into k segments and quantizing those segments independently using k sub-quantizers q_j where q_j is a quantizer of low complexity associated with j_{th} segment of the input vector.

We employ product quantization to compress raw descriptor $h = (h_1 = (h_1^1, ..., h_1^m), ..., h_k = (h_k^1, ..., h_k^m))$, where $n = m * k$ is the descriptor dimension. h is structured into k parts by location bins, each part having m dimensions. We quantize the vectors of the k parts independently using k sub quantizers. Sizes z of all sub-quantizers' quantization dictionaries are the same. In that case, the whole representation space is given by z^k.

A sub-quantizer q maps a m-dimensional vector h_j to $q(h_j) \in C = \{c_i, c_i \in R^m, i \in I\}$. $I = \{0, ..., z-1\}$. C is a set of reproduction values to represent original vectors. The quality of a quantizer is measured by the mean squared error(MSE) between the input vector and its reproduction value, $MSE(q) = E(q(h) - h)^2$. We greedily learn subquantizers by minimizing MSE using Lloyd's algorithm over raw descriptors extracted from training images. The compressed descriptor is represented by a short code composed of all sub-quantizers' indices.

Parameter z decides quantization error as well as the size of quantized descriptor. When we use fixed-length coding for the sub-quantizers' output, the bits needed for a descriptor is $k * log_2 z$.

2.4. Descriptor distance calculation

Since descriptor h captures the probability distribution of gradient magnitudes and orientations, we use Kullback-Leibler divergence to compare descriptors. The distance between two descriptors x and y is defined as $dist(x, y) = \sum_{l=1}^{k} KL(x_l, y_l)$ where $KL(x_l, y_l) = \sum_{d=1}^{m} (x_l^d \log \frac{x_l^d}{y_l^d} + y_l^d \log \frac{y_l^d}{x_l^d})$.

To accelerate distance computation, we approximate the exact distance of descriptors using quantized descriptors as $dist(q(x), q(y)) = \sum_{l=1}^{k} KL(q(x_l), q(y_l))$. Distance between each pair of reproduction values $q(x_l)$ and $q(y_l)$ is pre-computed and stored in a table. Thus the distance $dist(q(x), q(y))$ can be efficiently calculated with k look-up table and addition operations.

Table 1: Details of 5 experiment data sets

category	match pairs	non-match pairs	query image	reference image
mixed text+ graphics	3000	30,000	1500	1000
paintings	400	4000	400	100
frames	400	4000	400	100
landmarks	3805	48675	3514	11083
common objects	2550	25500	2550	7650

3. EXPERIMENTS

Following the evaluation framework of CDVS[10], we evaluate PQ-WGLOH in the context of image retrieval, pairwise image matching and object localization. We first compare PQ-WGLOH with SIFT, GLOH and state of the art low bit rate descriptor CHoG. Then the scalability of PQ-WGLOH is validated by continuously varying the upper bound of total bits allowed for an image.

3.1. Experiment data sets

Eight data sets contributed to CDVS[10], including ZuBud, UKY, Stanford, ETRI, PKU, Telecom Italia, Telecom SudParis and Huawei are used in experiments. There are 30256 images in total which are assigned to 5 categories by contents, i.e. mixed text and graphics, paintings, frames captured from video clips, landmarks and common objects. All data sets provide the annotation files of pairs of match and non-match images, query and reference image lists. In addition, mixed text and graphics also provide bounding boxes for each pair of match images for localization experiment. Refer to Table.1 for more details about the data sets.

3.2. Pairwise matching

Matching operation is performed for each pair of images listed in annotation files. We use KL divergence as distance measure for PQ-WGLOH as well as CHoG, and L2-distance for SIFT and GLOH. Feature descriptors across two images are compared using the ratio test [1] with a threshold of 0.6. Potential feature matches are then verified by geometric consistency checking with RANSAC.

3.2.1. *Comparison of SIFT, GLOH, CHoG and PQ-WGLOH*

We use the SIFT implementation provided by Vedaldi et al[12] where each descriptor occupies 1024 bits. CHoG descriptors are generated using the executable program published by Chandrasekhar[9]. One CHoG descriptor takes 72 bits. We implement our own GLOH which occupies 576 bits. For PQ-WGLOH, we set $k = 9$, $m = 8$ and $z = 256$. Thus each descriptor eats up 72 bits.

As can be seen in Fig.3, PQ-WGLOH performs slightly worse than SIFT and GLOH with 10 times fewer bits. Moreover PQ-WGLOH outperforms CHoG descriptors at the same bits by 20% in terms of true positive rate at a given false positive rate. The superiority of PQ-WGLOH over CHoG probably lies in the selection of reproduction values. For CHoG, reproduction value v_i of node i is obtained as $v_i = 2^{-b_i}$ where b_i is the depth of tree node i. PQ-WGLOH, however, learns reproduction values in a data-driven manner which minimizes MSE. Note that all sub-quantizers in our experiments are learnt from 6000 images collected from Flicker. Hence we may claim sub-quantizers actually captured general structures which are independent of test data.

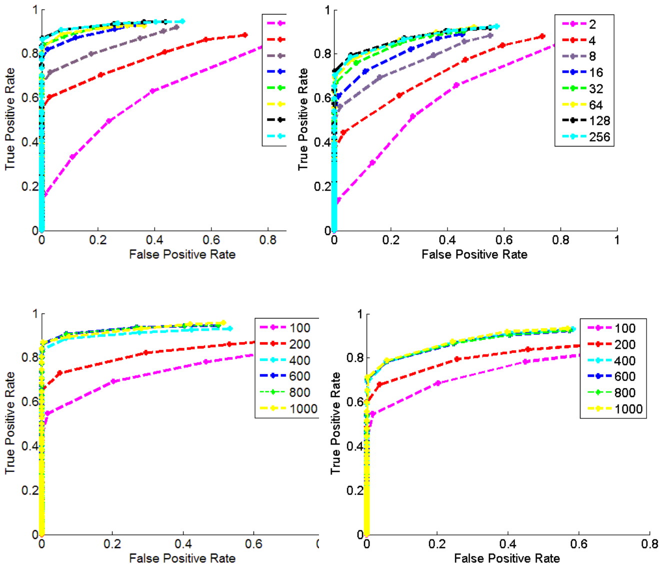


Fig. 5: ROC curve on PKU(left) and UKY(right)(1):Dictionary size z varies from $2 - 256$(top), (2): Descriptor number n varies from $100 - 1000$(bottom).

3.2.2. *Scalability*

We validate PQ-WGLOH's performance when the maximum bits allowed for an image vary. Total bits for representing an image are decided by the number of descriptors n and dictionary size z.

We first study the situations when z varies from 2 to 256. As shown in Fig.5(top), the performance decreases slightly when z varies from 256 to 16. When z is set between 8 and 16, matching performance drops by about 10% but is still acceptable. If $z < 8$, the performance decreases significantly.

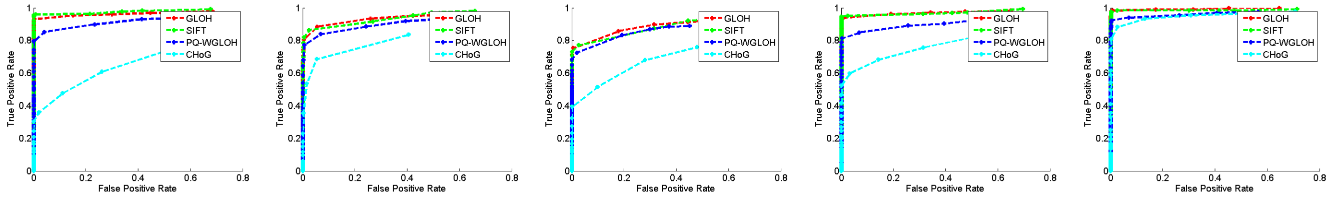
We then set upper bounds, ranging from 100 to 1000, of the number of selected descriptors for one image. Fig.5(bottom) shows the pairwise image matching result. We find that 400 descriptors are sufficient to produce satisfactory results. Moreover, performance decreases dramatically when fewer than 200 descriptor are used.

3.3. Object localization

Each query image is with a bounding quadrilateral covering corresponding regions of the reference image. The quadrilateral is specified by the coordinates of 4 vertexes.

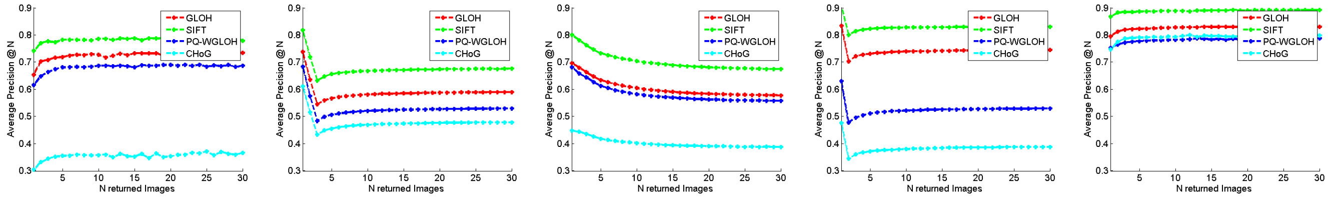
For a pair of query and reference images, we first match descriptors using ratio test. Based on locations of matched descriptors, the homography transformation from reference to query image is estimated. By the transformation, we project the four vertexes of reference quadrilaterals to query image. Localization accuracy is measured by a ratio: the overlapping area of back-projected and query quadrilaterals versus the total area filled by both quadrilaterals.

The overlap ratio for PQ-WGLOH is 0.8604, which is slightly worse than SIFT and GLOH, with 0.8902 and 0.8729 respectively. CHoG achieves localization accuracy of 0.8134.



(a) painting (b) common (c) landmark (d) mixed (e) frame

Fig. 3: ROC curve of pairwise matching on 5 categories of images.



(a) painting (b) common (c) landmark (d) mixed (e) frame

Fig. 4: AP @ N curve of image retrieval on 5 categories of images.

3.4. Image retrieval

Our retrieval pipeline applies the state of art inverted indexing techniques [13]. Instead of using K-means to construct visual vocabularies, we employ scalable vocabulary tree(SVT) method[14] to speed up vocabulary learning and descriptor quantization process. Branch factors and depth of SVT are set to 18 and 4 respectively, producing about 104976 leave nodes. Given a query, TF-IDF scheme is employed to score database images for retrieval. We adopt Average Precision @N for performance measure.

Fig.4 shows the retrieval results on 5 categories of data sets. SIFT and GLOH achieve about 10% higher AP @N than our PQ-WGLOH. The poor performance of PQ-WGLOH may be attributed to large quantization error from two-round quantization. CHoG suffers from the similar problem as PQ-WGLOH, but the AP @N is much lower than PQ-WGLOH on all data sets except video frames.

4. CONCLUSION

We proposed a compact yet discriminative local feature descriptor PQ-WGLOH with product quantization. The compactness reduces the load on wireless networks for visual search. Moreover, by precomputing a distance table, descriptor matching with low complexity is allowed. Extensive experiments on 8 data sets in the context of image retrieval, pairwise image matching and object localization show that PQ-WGLOH achieves comparable performance with SIFT and GLOH. In particular, PQ-WGLOH outperforms state of art CHoG.

5. ACKNOWLEDGEMENTS

This work was supported by the National Basic Research Program of China under contract no. 2009CB320902 and 2009CB320904, in part by grants from the Chinese National Nature Science Foundation under contract no. 60902057, and in part by the CADAL Project Program.

6. REFERENCES

[1] D.G. Lowe, “Distinctive image features from scale-invariant keypoints,” *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[2] H. Bay, T. Tuytelaars, and et al., “Surf: Speeded up robust features,” *ECCV*, pp. 404–417, 2006.

[3] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *PAMI*, pp. 1615–1630, 2005.

[4] M. Brown, G. Hua, and et al., “Discriminative learning of local image descriptors,” *PAMI*, pp. 43–57, 2010.

[5] Y. Ke and R. Sukthankar, “Pca-sift: A more distinctive representation for local image descriptors,” *CVPR*, 2004.

[6] A. Torralba, R. Fergus, and et al., “Small codes and large image databases for recognition,” *CVPR*, 2008.

[7] C. Strecha, A. Bronstein, and et al., “Ldahash: Improved matching with smaller descriptors,” *PAMI*, 2011.

[8] T. Tuytelaars and C. Schmid, “Vector quantizing feature space with a regular lattice,” *ICCV*, 2007.

[9] Chandrasekhar and el al., “Chog: Compressed histogram of gradients a low bit-rate feature descriptor,” *CVPR*, 2009.

[10] Yuri Reznik et al., “Call for proposals for compact descriptors for visual search,” 2011.

[11] R.M. Gray and D.L. Neuhoff, “Quantization,” *Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[12] A. Vedaldi and B. Fulkerson, “VLFeat: An open and portable library of computer vision algorithms,” http://www.vlfeat.org/.

[13] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” *ICCV*, 2003.

[14] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” *CVPR*, 2006.