Q-SIFT: EFFICIENT FEATURE DESCRIPTORS FOR DISTRIBUTED CAMERA CALIBRATION

Chao Yu, Gaurav Sharma

Electrical and Computer Engineering Dept.,
University of Rochester, Rochester NY 14627
{chyu,gsharma}@ece.rochester.edu

ABSTRACT

We consider camera self-calibration, i.e. the estimation of parameters for camera sensors, in the setting of a visual sensor network where the sensors are distributed and energy-constrained. With the objective of reducing the communication burden and thereby maximizing network lifetime, we propose an energy-efficient approach for self-calibration where feature points are extracted locally at the cameras and efficient descriptions for these features are transmitted to a central processor that performs the self-calibration. Specifically, in this work we use reduced-dimensionality quantized approximations as efficient feature descriptors. The effectiveness of the proposed technique is validated through feature matching, and epipolar geometry estimation which enable self-calibration of the network.

Index Terms— Local feature descriptor, visual (image) sensor network, self-calibration, energy constraint.

1. INTRODUCTION

Estimating the geometry of a camera network from the image contents only, i.e. *self-calibration* [1,2], is a critical foundation for many applications such as stereo, 3D modeling and tracking. In visual sensor networks (VSN) that consist of nodes formed by portable devices with imaging and communication capabilities, self-calibration should also take into account the energy consumption at these cameras, since these devices are usually battery-powered thus energy is usually the dominating limiting factor on the utility of these networks. The competing requirements in VSN applications (e.g. smart surveillance [3]) include long duration of unattended operation and limited energy supply, and motivate our investigation of an energy efficient approach for the self-calibration problem in a VSN.

We consider an application scenario illustrated in Fig. 1, where battery-powered image sensors are deployed to monitor a target region. Each of these cameras only communicates with a central processor (CP) which attempts self-calibration of the cameras in the network. We assume that the energy consumption at the camera nodes is dominated by the power required to send data to the CP, and thus aim at minimizing the data that these cameras need to transmit. For this purpose, the cameras locally extract feature points in the captured image, and only send the descriptions for these feature points to the CP. We use the scale-invariant transform (SIFT) based difference-of-Gaussian (DoG) features [4] and investigate efficient description of these features by using dimensionality reduction and quantization.

This work is supported in part by the National Science Foundation under grant number ECS-0428157.

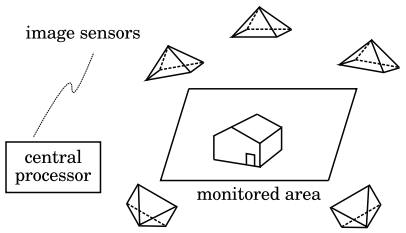


Fig. 1. N cameras deployed to monitor a 3D scene. A CP communicates with these cameras and coordinates the network.

Dimensionality reduction for description vectors of image features has been considered in prior work [5,6]. In the context of VSN, [5] also considers the selection of a subset of image features. However, these dimensionality-reduced features are still represented in a high resolution (32 bits in [5]). As we demonstrate, well-designed quantization for these descriptors reduces the communication cost dramatically. Experimental results show that a 20 dimensional feature descriptor using 20×4 bits, each dimension represented by 4 bit, sustains a distinctive description for a DoG feature. Another potential application of the efficient feature description is image retrieval [7], where low dimensional, distinctive features of an image are obtained and used to calculate the similarity of this image with an query image.

2. FEATURE EXTRACTION AND DESCRIPTION

We describe the features used in our approach and their efficient descriptions. These operations are performed by the cameras in the VSN.

Affine invariant feature detection [8] has been extensively studied recently. These features are robust to certain level of scale, illumination and viewpoint changes across multiple views. We use the DoG feature detector of SIFT [4], where feature points and the corresponding scales are selected from local extrema in the DoG pyramids. Orientation of the feature point is assigned to be the dominating gradient orientation in a neighborhood window. The coordinates, gradient orientations are rotated relative to this assigned orientation to obtain rotation-invariance. The descriptors are calculated as the normalized histogram of gradients around the selected feature point in the scale-space image. Using this process, each feature point is

represented by a 128-dimensional feature vector. Since the descriptor is normalized according to the orientation of the feature point and estimated in the scale-space, they sustain certain robustness under rotation and scaling. This feature descriptor, referred to as the scale-invariant feature transform (SIFT), performs better than other descriptors as shown in a comparative study [9].

The SIFT descriptor consists of 128 real numbers, which requires 128×32 bits if each real number is represented by 4 bytes. The sizes of the feature descriptors are usually comparable with the image itself, or even larger. In the application scenario we consider, an efficient feature description is desirable in order to reduce communication costs. To this end, we consider dimensionality reduced and quantized feature description, and investigate the trade-off between the efficiency of feature description and the distinctiveness of these descriptors.

2.1. Dimensionality Reduced Feature Description

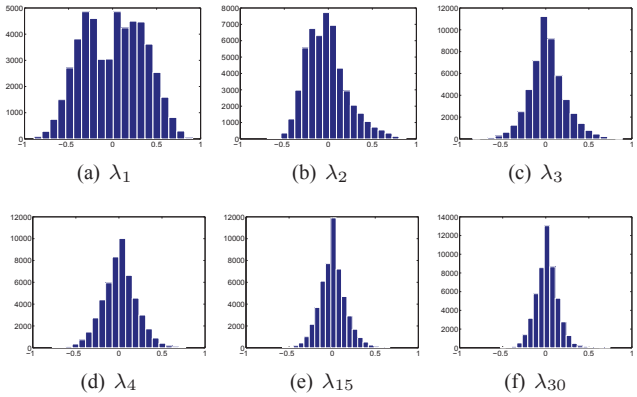
One (straightforward) approach to reduce the dimensionality of the SIFT features consists of performing a principal components analysis (PCA) on the 128×1 feature vectors computed by the SIFT algorithm [4]. For adequate performance, however, this "SIFT-PCA" approach [5] requires between $k = 40$ to 80 coefficients and additionally the $128 \times k$ describing the PCA transformation. Superior performance is obtained with an alternative approach called PCA-SIFT [6] that first performs the PCA on the rotationally aligned DoG gradient distributions at the feature points and then directly utilizes the PCA coefficients as the feature descriptors for feature matching. Specifically, for PCA-SIFT, using the 2×1 gradients in a 39×39 neighborhood window for each feature point and rotating these gradients with respect to the dominant orientation provides a $2 \times 39 \times 39 = 3042$ dimensional vector. It is observed that in this case, using $k = 36$ (or even $k = 20$) of the principal components provides a feature matching performance that is close to (or in some cases better than) the performance obtained with the original 128 dimensional SIFT descriptor. We thus adopt PCA-SIFT as the basis of our feature description and proceed to consider quantization of PCA-SIFT feature vectors.

2.2. Quantized Feature Description

The quantization for feature descriptors is desirable for both compact representation and efficient computation (matching), the former is important in the communication system we consider, the latter is critical for an image retrieval system.

We first perform an experimental study for the coefficients in a PCA-SIFT description. We use 24 images from the Kodak color image database[1], from which 51672 features points are detected and used as training data. The feature vector $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_k]$ is normalized so that $\|\boldsymbol{\lambda}\| = 1$. We thus obtain 51672 samples for each normalized coefficient λ_i, $i = 1, 2, \ldots, k$, in the k dimensional feature descriptors. Figure 2 shows the histograms for several of these coefficients. From these histograms, we see that except for the first two coefficients λ_1,λ_2, the remaining coefficients exhibit similar distributions (approximately Laplacian with nearly the same variances). Higher ordered coefficients (λ_{30}) are only slightly more concentrated at zero compared to lower ordered coefficients

[1]Images available at `http://www.site.uottawa.ca/~edubois/demosaicking/`



(a) λ_1 (b) λ_2 (c) λ_3

(d) λ_4 (e) λ_{15} (f) λ_{30}

Fig. 2. Histogram of selected coefficients of the PCA feature description, obtained from training data using 51672 feature points detected in the Kodak image database.

(λ_3). This observation suggests that more bits should be allocated for lower ordered coefficients. Since the differences are small, however, we allocate an equal number of bits to each of the coefficients, which we denote by t.

For each coefficient we utilize a non-uniform scalar quantizer[2] obtained by using the K-means [10] algorithm, which in the scalar scenario in consideration corresponds to the Lloyd-Max [11] algorithm with the empirical histogram as the probability distribution function. For each coefficient λ_i, this process yields $N = 2^t$ representation points $\{c_{ij}\}_{j=1}^{N}$, for $i = 1, 2, \ldots, k$ and corresponding intervals $\{(l_{ij}, u_{ij})\}_{j=1}^{N}$, where the interval (l_{ij}, u_{ij}) is represented by c_{ij} in the quantized representation. By building up this quantization table represented by $\{(l_{ij}, u_{ij}; c_{ij})\}_{j=1}^{N}$, for each coefficient λ_i, a new feature description can be represented by $kt = k \times \log_2 N$ bits. We consider different values of $N = 2, 4, \ldots, 256$ to examine the effect of different quantization levels.

For a typical setting of $k = 20, N = 16$, and assuming each of the image coordinates is represented by 4 bytes, the description for a feature point can be represented by 18 bytes. If 500 feature points are detected and described, the total message length is 9KB (kilo bytes), which is significantly smaller than the approach in [5], where a typical message length is 120KB and contains information for fewer features.

3. ESTIMATING EPIPOLAR GEOMETRY VIA ROBUST FEATURE MATCHING

We investigate the robustness of the quantized feature descriptors by considering the estimation of epipolar geometry between two views of a scene. The epipolar geometry is described by the *fundamental matrix* $\mathbf{F} \in \mathbb{R}^{3\times3}$, for which the determinant $\det(\mathbf{F}) = 0$. For a point $\mathbf{p}_i$ in image $\mathcal{I}_1$, the correspondence $\mathbf{q}_j$ in image $\mathcal{I}_2$ lies on the line defined by $\mathbf{l}_i = \mathbf{F}\mathbf{p}_i$, i.e. $\mathbf{q}_j^T \mathbf{F} \mathbf{p}_i = 0$, which is also referred to as the *epipolar constraint* [1]. Estimating the fundamental matrix is the critical step in many self-calibration methods [1] and the

[2]Note the PCA decorrelates the coefficients, so per coefficient scalar quantization causes limited performance degradation over vector quantization.

accuracy of fundamental matrix estimation therefore allows us to assess the performance of our efficient feature description scheme in an energy-efficient approach for the self-calibration of a VSN.

Tentative matches between features in images $\mathcal{I}_1$ and $\mathcal{I}_2$ are first identified. For feature $\mathbf{p}_i$ in $\mathcal{I}_1$, the feature $\mathbf{q}_j$ in image $\mathcal{I}_2$ is identified as a tentative match if

$$\mathsf{D}(\boldsymbol{\lambda}(\mathbf{p}_\mathsf{i}), \boldsymbol{\lambda}(\mathbf{q}_\mathsf{j})) < \rho \mathsf{D}(\boldsymbol{\lambda}(\mathbf{p}_\mathsf{i}), \boldsymbol{\lambda}(\mathbf{q}_\mathsf{l})), \text{for } \mathsf{l} \neq \mathsf{j}, \qquad (1)$$

where $\boldsymbol{\lambda}(\mathbf{p}_i)$ represents the the description vector of $\mathbf{p}_i$, $\mathsf{D}(\cdot)$ is a distance function (in our case is the Euclidean distance) and ρ is a parameter which defines the matching threshold. Smaller values of ρ result in more reliable matches, however the number of matches decrease.

The estimation of fundamental matrix is quite sensitive to noise and false matches. The Random Sample Consensus (RANSAC) algorithm [12] is used to eliminate false matches. The RANSAC algorithm operates in an iterative manner. In each iteration, 8 tentative correspondences are randomly selected to estimate the fundamental matrix using a linear least-squares approach [1] which is derived from the epipolar constraint. In addition, we normalize the input data and impose the rank-two constraint for better estimation accuracy. Next, from the estimated fundamental matrix, the epipolar line $\mathbf{l}_i$ is calculated for each point $\mathbf{p}_i$. Inlier (correct) matches are identified as the tentative matches for which the distance $\mathbf{q}_j^T \mathbf{l}_i$ of the tentative match $\mathbf{q}_j$ from the epipolar line lies below a pre-determined threshold. The iteration ends when the average distances of the inlier matches with the epipolar line is smaller than a threshold or a maximum number of iterations have been executed.

4. EXPERIMENTAL RESULTS

We examine the performance of the proposed feature description by estimating the epipolar geometry of several image pairs under various imaging situations. Four image pairs shown in Fig. 4 are used in our experiments[3] to represent images under viewpoint change, scaling and rotation, blurring, and illumination change, respectively.

We compare several feature descriptions: the original SIFT, PCA-SIFT, and Q-SIFT proposed in this paper. We consider several quantization schemes with $t = 2, 4, 8$ representing the number of bits used to represent each feature coefficient. For PCA-SIFT and Q-SIFT, we also investigate the effect of the dimensions of feature descriptions by considering $k = 10, 20, 36$. Using feature descriptions extracted from each pair of images, the procedure described in Section 3 is used to estimate matched features and the epipolar geometry. The ratio in (1) is set as $\rho = 0.85$. Experimental results are summarized in Table 1, indicating several interesting observations:

i) SIFT is more robust under viewpoint change. This is indicated by experiments with Graffiti, where the largest number of inlier matches are identified using SIFT. PCA-SIFT and Q-SIFT also achieve robust estimates of the epipolar geometry when $k = 20, 36$. However, insufficient inlier matches are identified when $k = 10$.

ii) In other experiments with image pairs under rotation, scaling, blurring and lighting change, the epipolar geometry is successfully estimated and large number of inlier matches are found. Furthermore, the number of inlier matches increase as the description length k increases. For instance, experiments for the boat image using Q-SIFT with $t = 2$, yielded $310, 496, 549$ inlier matches, for $k = 10, 20, 36$, respectively. The residues in these experiments, i.e. the mean absolute difference between the inlier matches and the corresponding epipolar lines are obtained as $0.273, 0.242, 0.295$, which are close to the residue for the original SIFT algorithm (0.215).

iii) Feature descriptions using $t = 4, 8$ bits for each coefficient result in significant larger number of inlier matches compared to $t = 2$, especially under viewpoint change. However, the experiment results with $t = 8$ bits offer only a slight performance improvement over $t = 4$ bits. For instance, experiments for the graffiti image using Q-SIFT with $k = 20$, yielded $86, 130, 145$ inlier matches for $t = 2, 4, 8$, respectively. Another experiment with the boat image using Q-SIFT with $k = 20$, obtained $496, 578, 598$ inlier matches, respectively, for $t = 2, 4, 8$.

Given the observations above, Q-SIFT with $k = 20, t = 4$ clearly achieves an efficient trade-off between compact representation and distinctive feature description. Combined with the 4-bytes per image coordinate (for capturing feature locations with sub-pixel accuracy), the resulting representation requires 18 bytes for the description of each feature point, which is a significant improvement over a previously proposed alternative [5].

We also visualize the epipolar geometry estimated for the Graffiti image pair using Q-SIFT with $k = 20, t = 4$. As illustrated in Fig. 4, randomly selected 10 inlier matches are connected across two views, the epipolar line for these 10 features are also shown. As expected, these epipolar lines intersect at the *epipole* [1].

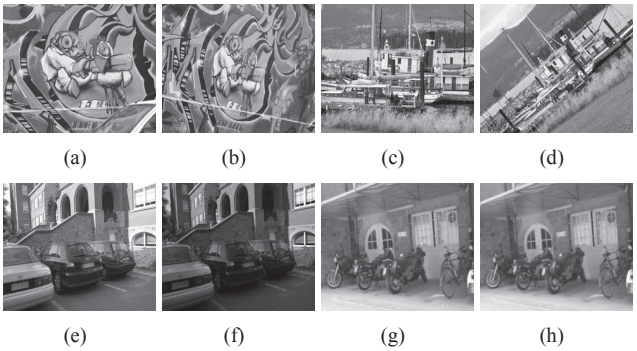


Fig. 3. Test images used in experiments. Different imaging situations are represented by: (a,b) Graffiti, viewpoint change. (c,d) boat, rotation and scaling. (e,f) car, illumination change. (g,h) bike, blurring.

5. DISCUSSION AND CONCLUSION

In the context of energy-efficient self-calibration of a VSN, we propose an efficient description for affine-invariant features in an image. The feature descriptions are locally computed and transmitted to a CP for the self-calibration. Each feature point is represented by 18 bytes, thus a typical message length is 9KB if 500 features are represented. The proposed technique, thus significantly outperforms the simple approach of sending a compressed image. Experimental results demonstrate that using a 20 dimensional feature vector, with each dimension represented by 4 bits, can achieve distinctive feature descriptions and an accurate estimate for the epipolar geometry. We

[3]These images are available from `http://www.robots.ox.ac.uk/~vgg/research/affine/`.

descriptor		graffiti			boat			bike			car		
		n_t	n_f	d_f	n_t	n_f	d_f	n_t	n_f	d_f	n_t	n_f	d_f
SIFT		365	251	0.368	730	663	0.215	602	559	0.255	376	355	0.219
PCA-SIFT	k=36	247	163	0.389	660	599	0.242	555	517	0.194	361	337	0.208
	k=20	260	140	0.334	674	582	0.247	559	512	0.247	365	340	0.170
	k=10	291	62	0.331	688	532	0.236	557	490	0.195	353	319	0.216
Q-SIFT (2)	k=36	168	87	0.376	601	549	0.273	521	481	0.212	340	323	0.250
	k=20	220	86	0.310	596	496	0.242	517	459	0.194	325	299	0.245
	k=10	318	8	0.271	580	310	0.295	463	337	0.263	280	217	0.234
Q-SIFT (4)	k=36	265	148	0.372	670	595	0.266	550	509	0.211	353	335	0.222
	k=20	281	130	0.460	684	578	0.215	571	508	0.273	359	337	0.183
	k=10	325	52	0.459	671	495	0.242	560	473	0.197	339	305	0.258
Q-SIFT (8)	k=36	265	143	0.353	666	598	0.213	559	517	0.176	363	340	0.218
	k=20	289	145	0.376	695	591	0.274	569	516	0.198	367	344	0.219
	k=10	317	51	0.376	701	529	0.256	573	488	0.203	360	323	0.241

Table 1. Estimation of epipolar geometry using different feature descriptors. The estimation result is described by three numbers n_t, n_f and d_f indicating, respectively, the number of initial tentative matches identified from feature description using the criterion (1), the number of final inlier matches satisfying the epipolar constraint, and the average distance of these final inlier matches to the corresponding epipolar line. For instance, in the experiment with graffiti using SIFT description, we obtain 365 initial matches, 251 inlier matches after estimating the epipolar constraint and the average distance to the epipolar line is 0.368. Q-SIFT (2) indicates using Q-SIFT and $t = 2$ bits for each feature coefficient. The results of using Q-SIFT with $t = 1$ is not satisfactory, and not included.

(a) (b)

Fig. 4. (a) 10 out of the 130 inlier matches in Graffiti. Corresponding features are connected by the line across two images. (b) The epipolar lines of the 10 features in the top image are plotted, the corresponding feature in the bottom image lies in this line. These epipolar lines intersect at the epipole.

conclude with a few remarks and our future work.

i) This paper focuses on efficient representation of a feature point. In order to achieve effective representation of an image, a method to select a subset of feature point from an image is required, such as in [5].

ii) Estimation of the epipolar geometry and robust matching serve as the first stage for self-calibration of a camera network, in continuing work we are exploring the application of the technique in the full self-calibration process and will be investigating the trade-off between efficient local feature representation and the accuracy of the estimated network geometry.

6. REFERENCES

[1] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. New York, NY, USA: Cambridge University Press, 2000.

[2] M. Pollefeys, R. Koch, and L. V. Gool, "Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters," *Intl. J. Computer Vision*, vol. 32, no. 1, pp. 7–25, 1999.

[3] A. Keshavarz, A. Tabar, and H. Aghajan, "Distributed vision-based reasoning for smart home care," *ACM SenSys Workshop on Distributed Smart Cameras (DSC06)*, 2006.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[5] Z. Cheng, D. Devarajan, and R. Radke, "Determining vision graphs for distributed camera networks using feature digests," *EURASIP Journ. Adv. in Sig. Proc.*, 2007.

[6] Y. Ke and R. Sukthankar, "PCA-SIFT: A More Distinctive Representation for Local Image Descriptors," in *IEEE Intl. Conf. Comp. Vision, and Pattern Recog.*, vol. 2, 2004.

[7] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Anal. Mach. Intel.*, pp. 1349–1380, 2000.

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Gool, "A Comparison of Affine Region Detectors," *Intl. J. Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.

[9] K. Mikolajczyk and C. Schmid, "A Performance Evaluation of Local Descriptors," *IEEE Trans. Pattern Anal. Mach. Intel.*, pp. 1615–1630, 2005.

[10] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[11] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Info. Theory*, vol. 28, no. 3, pp. 129–137, Mar. 1982.

[12] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. Assoc. Comp. Mach.*, vol. 24, no. 6, pp. 381–395, 1981.