PRUNING SIFT & SURF FOR EFFICIENT CLUSTERING OF NEAR-DUPLICATE IMAGES

Tushar Shankar Shinde, Anil Kumar Tiwari

Department of Electrical Engineering, Indian Institute of Technology Jodhpur, India
Email: {shinde.1, akt}@iitj.ac.in

ABSTRACT

Clustering and categorization of similar images using SIFT and SURF require a high computational cost. In this paper, a simple approach to reduce the cardinality of keypoint set and prune the dimension of SIFT and SURF feature descriptors for efficient image clustering is proposed. For this purpose, sparsely spaced (uniformly distributed) important keypoints are chosen. In addition, multiple reduced dimensional variants of SIFT and SURF descriptors are presented. Moreover, clustering time complexity is also improved by proposed contextual bag-of-features approach for partitioned keypoint set. The F-measure statistic is used to evaluate clustering performance on a California-ND dataset containing near-duplicate images. Clustering accuracy of the proposed pruned SIFT and SURF is found to be at par with traditional SIFT and SURF with a significant reduction in computational cost.

Index Terms— Image Clustering, SIFT, SURF, Bag-of-Features, Dimensionality Reduction

1. INTRODUCTION

In today's world where millions of images are flooded on the web every day, organizing and categorizing them is of paramount importance. According to a report [1], Facebook has around 100 billion images and the number is increasing exponentially. In image clustering, images are grouped in such a way that each cluster contains more alike images [2]. In general, image clustering involves the following steps: extracting features from the image, organizing them, and then classifying the image to specific cluster [3]. Many feature extraction algorithms are available in literature [4]. The Scale Invariant Feature Transform (SIFT) is one of the most popular feature extraction algorithm due to its invariant nature to scaling, rotation, translation, illumination changes, and small distortions [5]. SIFT and the several times faster version of SIFT named Speeded-Up Robust Features (SURF) [6] has been successfully used in various computer vision and image processing applications such as object recognition, panorama stitching, $3D$ modeling, robot localization and video tracking [7, 8].

This work was supported by the Visvesvaraya Ph.D. scheme for Electronics and IT Research Fellowship (MeitY, India).

Deep adaptive learning-based methods [9] could be computationally demanding over SIFT-based methods for clustering near-duplicate (ND) images. Moreover, SIFT-based clustering requires a substantial amount of computations mainly due to a large number of keypoints, and high dimensional feature descriptor. The non-maximal suppression algorithm enhanced the recognition of spatially distributed features [10]. Moreover, the cardinality of the keypoint set can be reduced by raising the keypoint inclusion threshold value [11]. However, this method produces an unstable and non-uniform number of keypoints. Many researchers have reduced the dimension of the SIFT descriptor [12, 13]. The SIFT variant that uses principal component analysis (PCA) for dimensionality reduction is PCA-SIFT. PCA-SIFT features are more compact than standard SIFT features. However, additional computations are required for applying PCA. The SIFT and SURF have been extensively used in visual word dictionary generation. The orderless bag-of-features (BoF) is found to be effective in image matching task [14]. However, the ordered BoF is observed to be more effective than the traditional BoF model [15]. The partitioned keypoint set dictionary generation is seldom investigated in the literature.

In this paper image clustering with a simple approach of pruning SIFT and SURF features are studied. The proposed work focused on three main aspects: (1) non-uniformity of keypoints, (2) larger dimension of the feature descriptor, and (3) partitioned keypoint set dictionary generation. For this, a technique for selection of uniformly distributed keypoints and reducing the dimension of the descriptor is presented such that the distinctiveness of SIFT and SURF is preserved. Moreover, the keypoint set is partitioned into multiple groups based on the strength of the features for dictionary generation. The clustering accuracy is evaluated using the F-measure statistic. The pruned SIFT and SURF provided encouraging results with significantly reduced computational cost.

The remainder of this paper is structured as follows: In Section 2, SIFT, and SURF based Image clustering approach is described in brief. In Section 3, an efficient clustering scheme based on pruned features is presented. Section 4 explains evaluation methodology. The experimental results which demonstrate the effectiveness of the proposed approach is discussed in Section 5. Section 6 concludes the paper.

2. SIFT, SURF, AND IMAGE CLUSTERING

SIFT algorithm [5] consists of four key steps: (1) scale-space extrema detection, (2) keypoint localization, (3) orientation assignment, and (4) keypoint descriptor representation. The 16×16 region around each keypoint is selected and divided into sixteen 4×4 sub-regions. Then for each sub-region, orientation histograms are computed with 8 bins (45° angles) each. The resulting feature descriptor of length 128 is normalized to unit length to introduce the invariance to changes in illumination.

However, SURF algorithm [6] consists of only two key steps: (1) keypoint detection, and (2) keypoint descriptor representation. The squared region around each keypoint is selected and divided into sixteen 4×4 sub-regions. Then, Haar wavelets are computed for each sub-region yielding 4 values. Hence, the SURF feature descriptor of length 64 is obtained.

Bag-of-Features (BoF) is a widely adopted visual feature descriptor of an image used for classification [16]. The SIFT or SURF features for all the test images (N) are extracted and the same is used for image clustering using the traditional BoF model. In general, the K-means clustering algorithm is used in BoF. In BoF [17], feature descriptors are clustered for the desired amount of visual words (K). Then considering each visual word as a bin, a histogram with K bins is obtained for each image. Next, a normalized histogram of size ($N \times K$) is used to cluster the test collection into a required number of image clusters X.

K-means is the most famous clustering algorithm in literature [18]. It is also referred to as Lloyd's algorithm [19]. In the computer science community, Lloyd's algorithm is widely used for generating visual dictionaries [20]. The running time complexity of Llyod's algorithm is $O(ndki)$, where n denotes the number of d-dimensional keypoints, k is the number of visual words, and i represents a number of iterations required until convergence. For clustering images in large datasets, the computational time required is remarkably high. Hence, a scheme to reduce n, d, and k is illustrated in this work.

3. PROPOSED SCHEME

The computational time for clustering images could be significantly reduced by reducing a total number of keypoints (n), reducing dimension (d) of the feature descriptors, and reducing the number of visual words (k). Significant reduction in visual word count could severely affect clustering performance and hence it is not preferred. However, keypoints could be partitioned into contextual bins and dictionaries could be generated for each bin in parallel, resulting in a reduction in visual word count associated with each bin. The proposed approach for simple keypoint reduction, the reduced dimensional variants of SIFT and SURF descriptors, and partitioned contextual dictionary generation is described in this section.

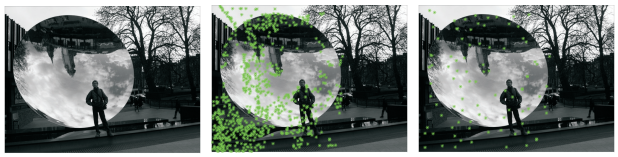

Fig. 1: (a) Original Image, (b) Actual keypoint set ($n = 1382$) (c) Uniformly distributed reduced keypoint set ($n_r = 125$).

3.1. Reducing Total Number of Keypoints

In general, the cardinality of keypoint set is in the order of 10^3 to 10^4 (depends on image size and content) [11]. For fast feature matching, it is desirable to restrict the maximum number of keypoints for each image. It is observed that the SIFT keypoints are densely spaced and an excessive number of keypoints provides only marginal enhancement in matching performance to justify the high computational cost. In effect, a smaller set of uniformly distributed keypoints is chosen. The keypoints are suppressed based on the keypoint strength (lower σ value represents lower strength), and only those having maximum strength in the predefined region are retained. To illustrate this, let us consider an image of resolution $H \times W$, with n and n_r denoting the actual and reduced number of keypoints respectively. For pruning keypoint set, the image containing multiple keypoints is divided into a rectangular grid of size $h \times w$. Real-valued grid size parameters h and w are calculated using (1).

$$h = \frac{H}{s}, \qquad w = \frac{W}{s}, \qquad s = \lfloor \sqrt{n_r} \rfloor \tag{1}$$

For each rectangular grid, only one keypoint with the highest strength (σ value) is chosen. This process not only resulted in the uniform distribution of keypoints but also significantly pruned a cardinality of keypoint set. The example of SIFT keypoint pruning is shown in Fig. 1.

3.2. Reducing SIFT and SURF Descriptor Dimension

Fig. 2 (a) depicts the standard 128 dimensional ($128D$) SIFT descriptor that distinctively describes the keypoint. To reduce the dimension of SIFT descriptor, the 4×4 sub-regions are combined such that, orientation histogram bins in the same directions are accumulated. The different combinations of combining these sub-regions resulted in different variants of SIFT descriptor. The three variants of the descriptor with different dimensions: $\{64D, 32D, 8D\}$ are shown in Fig. 2 (b) to (d). The procedure to reduce the SIFT descriptor dimension is illustrated in Table 1.

For example, $8D$ descriptor (Fig. 2 (d)) is formed by accumulating orientation histogram bins as follows. Here, each one of the sixteen sub-region is represented by total 8 orientations. Let $d = \{d_1, d_2, d_3, .., d_{128}\}$ be the standard 128D

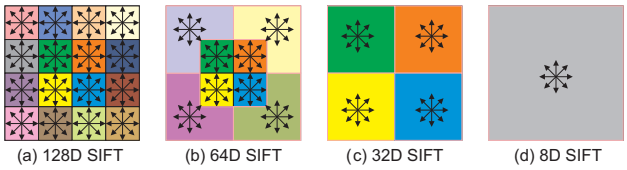

Fig. 2: (a) Standard $128D$ SIFT, (b) $64D$, (c) $32D$, (d) $8D$.

Table 1: Description of three variants of SIFT descriptor.

SIFT	Description
$64D$	Combine corner and neighboring boundary sub-regions ($8 \times 4 = 32D$), and retain central 2×2 sub-regions ($8 \times 4 = 32D$)
$32D$	Combine corner and neighboring sub-regions ($8 \times 4 = 32D$)
$8D$	Combine all sub-regions ($8D$), refer Fig. 2 (d)

SIFT descriptor such that $\{d_{i*8-7}, d_{i*8-6}, d_{i*8-5}.., d_{i*8}\}$ denote the orientations corresponding to the i^{th} sub-region. Then, the compact $8D$ feature descriptor $c = \{c_1, c_2, .., c_8\}$ is obtained as follows:

$$c = \left\{ \sum_{i=1}^{16} d_{i*8-7}, \sum_{i=1}^{16} d_{i*8-6}, .., \sum_{i=1}^{16} d_{i*8} \right\} \qquad (2)$$

The similar dimensionality reduction scheme is applied to SURF descriptor. SURF has only 4 values corresponding to each sub-region unlike SIFT, which has 8 orientation bins. The three variants of SURF descriptor with different dimensions: $\{32D, 16D, 4D\}$ are obtained by accumulating sub-regions in a similar way as shown in Fig. 2 (b) to (d).

Due to the reduction in the dimension of SIFT and SURF descriptor in this way, the local features tend to be more global while they are highly distinctive yet. The distinctiveness of the descriptor is preserved by accumulating orientation bins corresponding to the same directions.

3.3. Partitioned Contextual BoF Approach

The main objective of the BoF approach is to form a dictionary of visual feature descriptors. The keypoint set for a particular image is partitioned into predefined subsets based on the strength (σ values) of the keypoints (higher σ value indicates higher strength). Let keypoint set u be partitioned into j subsets such that each subset contains an equal number of keypoints. Then the more appropriate way of dictionary formation is to form separate dictionaries for each subset and combine those to obtain the overall dictionary. The conceptual illustration of the proposed scheme is shown in Fig. 3.

4. EVALUATION METHODOLOGY

4.1. Data Sets

For experiments, we generated $N = 3000$ test images. For this, $X = 30$ different classes with each class containing 5 near-duplicate images (with slightly varying viewpoints) resulting in total 150 photos are chosen from California-ND dataset [21]. Then, 20 alterations are applied to each of them, resulting in a total of 100 images in each class. All the test images are first converted to gray-scale before applying alterations. The list of alterations similar to that of the works in [22] is described in Table 2. Each image in the dataset has a spatial resolution of 1024×768 or 768×1024 pixels.

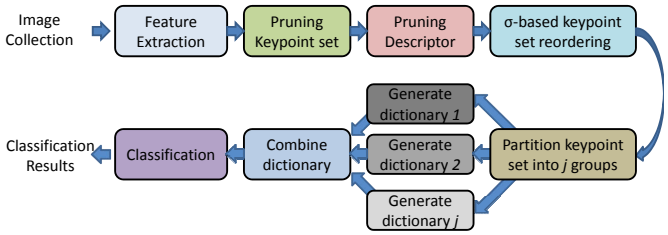

Fig. 3: A conceptual illustration of proposed scheme.

Table 2: Type of alterations applied to each image. The number in the parenthesis represents a total number of images generated for each type of alteration.

Alteration	Description
scale	scale by $\times 0.5, \times 0.75, \times 2$ (3)
rotate	rotate by 90°, 180°, 270° about its center (3)
crop	crop 90%, 75%, preserve center region (2)
intensity	change intensity by ± 25, ± 50 units (4)
blur	apply Gaussian blurring by $\sigma = \{1, 2, 5\}$ (3)
noise	add 1% salt and pepper noise (1)
rotate+crop	rotate by 180° about its center and crop 75%, preserve center region (1)
scale+rotate	scale by $\times 0.75$ and rotate by 180° about its center (1)
scale+crop	scale by $\times 2$ and crop 75%, preserve center region (1)

4.2. Performance Measure Metric

F-measure is employed to empirically demonstrate the effectiveness of the proposed scheme for image clustering. The mean $F1$ score is extensively used to measure clustering accuracy using two statistical parameters: precision (p) and recall (r). Precision and recall, computed for each class, are the ratio of true positives (tp) to all predicted positives (tp+fp) and ratio of true positives (tp) to all actual positives (tp+fn), respectively, as in (3).

$$p = \frac{tp}{tp + fp}, \qquad r = \frac{tp}{tp + fn} \qquad (3)$$

where fp and fn represents false positives and false negatives, respectively. The $F1$ score is calculated for each class and average of these $F1$ scores resulted in mean $F1$ score ($mF1$) as in (4).

$$F1 = 2\frac{pr}{p + r}, \qquad mF1 = mean(F1) \qquad (4)$$

The $F1$ score weights precision and recall equally. The score $mF1 \in [0, 1]$, with 1 representing perfect clustering.

For computational complexity analysis, speed-up over traditional SIFT is computed for all pruned combinations and it is computed as:

$$Speedup = \frac{Computational\ Time_{traditional\_SIFT}}{Computational\ Time_{pruned\_approach}} \quad (5)$$

All the experiments are implemented in MATLAB 2015a running on 64-bit Windows 7 platform with Intel Xeon(R) CPU E3-1215 v5 @ 3.30 GHz with 8.0 GB RAM.

5. EXPERIMENTAL RESULTS

The simulation results for different combinations of keypoint set, descriptor dimension, and partitioned contextual dictionary generation are illustrated in Table 3. For experiments, the visual word count of $K = 300$ is empirically chosen. The support vector machine (SVM) is used for multiclass classification trained using the one-versus-all rule. The parameters of SVM are empirically set and kept fixed for all the tests.

It is observed in Table 3 that reducing cardinality of the keypoint set does not significantly affect clustering accuracy. Among $\{25, 50, 100, 200, All\}$ cardinalities of the keypoint set, the cardinality of 100 provided a trade-off between computational cost and clustering accuracy. Experiments on the benchmark datasets show that selecting distributed keypoints in this way, as compared to selecting keypoints based on global maximum strength, provides better accuracy.

Among all SIFT variants, standard $128D$ SIFT, as expected, outperformed others. The $64D$ SIFT descriptor achieved clustering accuracy at par with $128D$ SIFT and outperformed traditional $64D$ SURF. It is also observed that a further decrease in the descriptor dimension resulted in a marginal reduction in accuracy, but significant speed-up is achieved.

We have compared results for traditionally adopted non-partitioned dictionary generation against proposed partitioned contextual BoF approaches. The results for no, 2, and 3 partitions, illustrated in Table 3 depicts that partitioned dictionary generation provided clustering performance at-par with traditional BoF approach, but at much higher speed.

The standard $128D$-SIFT and $64D$-SURF is used to study the effect of pruning keypoint set, whereas the effect of reduction in descriptor dimension is analyzed at fixed cardinality of keypoint set $n_r = 100$. Fig. 4 clearly depicts that the clustering performance is least affected when the keypoint set is reduced as compared to the effect of reduction in the descriptor dimension. Moreover, the partitioned dictionary resulted in inferior performance due to sub-optimal visual word creation. However, employing all the three techniques simultaneously resulted in 121 thousand times speed-up but provided only 86% accuracy. Hence, the appropriate combination should be chosen based on the accuracy versus speed trade-off. For the very tight computational budget, selection of at least 25 number of keypoints and use of $8D$ SIFT descriptor is recommended. The similar results are obtained for SURF as well.

Table 3: SIFT Vs SURF Clustering Performance.

			No Partition		2 Partitions		3 Partitions	
	D	N	mF1	Speedup	mF1	Speedup	mF1	Speedup
SIFT	128D	All	0.9956	1	0.9814	4.1	0.9644	12.7
	128D	200	0.9933	69	0.9811	388	0.9629	802
	128D	100	0.9933	279	0.9793	1238	0.9600	2963
	128D	50	0.9933	1197	0.9789	4522	0.9522	12866
	128D	25	0.9915	3908	0.9763	3908	0.9311	39477
	64D	100	0.9900	453	0.9758	2117	0.9589	6065
	32D	100	0.9833	572	0.9742	3634	0.9489	7962
	8D	100	0.9800	1063	0.9689	5003	0.9256	12771
	8D	50	0.9789	3337	0.9456	17052	0.8867	39633
	8D	25	0.9767	11526	0.9386	52817	0.8600	121868
SURF	64D	All	0.9861	2.3	0.9728	8.9	0.9546	27.4
	64D	200	0.9823	132	0.9701	388	0.9521	1723
	64D	100	0.9811	567	0.9698	1238	0.9515	6095
	64D	50	0.9807	2402	0.9683	4522	0.9507	27602
	64D	25	0.9789	8132	0.9672	3908	0.9492	87497
	32D	100	0.9800	953	0.9690	2117	0.9510	13052
	16D	100	0.9733	1123	0.9611	3634	0.9438	19042
	4D	100	0.9200	2274	0.8964	5003	0.8216	28438

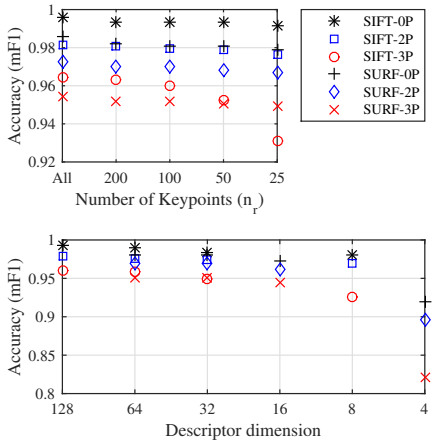

Fig. 4: SIFT Vs SURF performance comparison where 0P, 2P, and 3P corresponds to No, 2, and 3 partitions respectively.

6. CONCLUSION

A simple technique to reduce cardinality of keypoint set by selecting uniformly distributed features and three variants of reduced dimensional SIFT and SURF descriptor is presented. Additionally, partitioned contextual dictionary generation is proposed to further reduce the computational complexity of visual word dictionary generation. Experimental results show that pruning keypoint set to only 1% of the original keypoint set provides less than 0.5% decrease in the clustering performance. Moreover, reducing the dimension of the descriptor to only $8D$ of the traditional SIFT descriptor provides less than 1.5% decrease in the clustering performance. It is noted that pruning SIFT and SURF does not significantly affect the discriminating capability of SIFT and SURF features. The future work would be to evaluate the proposed technique for a larger set of near-duplicate and normal image collection.

7. REFERENCES

[1] Infosys, "Future of image technologies in financial services," *Retrieved 3rd January*, 2017.

[2] Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueting Zhuang, "Image clustering using local discriminant models and global integration," *IEEE Transactions on Image Processing*, vol. 19, no. 10, pp. 2761–2773, 2010.

[3] Huan Liu and Lei Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[4] Krystian Mikolajczyk and Cordelia Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

[5] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[7] David G Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Ieee, 1999, vol. 2, pp. 1150–1157.

[8] Matthew Brown and David G Lowe, "Automatic panoramic image stitching using invariant features," *International journal of computer vision*, vol. 74, no. 1, pp. 59–73, 2007.

[9] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Deep adaptive image clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5879–5887.

[10] Ran Song and John Szymanski, "Well-distributed sift features," *Electronics letters*, vol. 45, no. 6, pp. 308–310, 2009.

[11] Jun Jie Foo and Ranjan Sinha, "Pruning sift for scalable near-duplicate image matching," in *Proceedings of the eighteenth conference on Australasian database-Volume 63*. Australian Computer Society, Inc., 2007, pp. 63–71.

[12] Yan Ke and Rahul Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*. IEEE, 2004, vol. 2, pp. II–II.

[13] Nabeel Younus Khan, Brendan McCane, and Geoff Wyvill, "Sift and surf performance evaluation against various image deformations on benchmark dataset," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*. IEEE, 2011, pp. 501–506.

[14] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. IEEE, 2006, vol. 2, pp. 2169–2178.

[15] Yang Cao, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang, "Spatial-bag-of-features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3352–3359.

[16] Xiaoli Yuan, Jing Yu, Zengchang Qin, and Tao Wan, "A sift-lbp image retrieval model based on bag of features," in *IEEE international conference on image processing*, 2011.

[17] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 494–501.

[18] Anil K Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[19] Stuart Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[20] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[21] Amornched Jinda-Apiraksa, Vassilios Vonikakis, and Stefan Winkler, "California-nd: An annotated dataset for near-duplicate detection in personal photo collections," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*. IEEE, 2013, pp. 142–147.

[22] Jun Jie Foo, Justin Zobel, and Ranjan Sinha, "Clustering near-duplicate images in large collections," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*. ACM, 2007, pp. 21–30.