Performance Evaluation of State-of-the-art Filtering Criteria Applied to SIFT Features

Silvère Konlambigue, Paul Honeine, Jean-Baptiste Pothin, Abdelaziz Bensrhair

Performance Evaluation of State-of-the-Art Filtering Criteria Applied to SIFT Features

Silvère KONLAMBIGUE[⋆†], Jean-Baptiste POTHIN[⋆], Paul HONEINE[†], Abdelaziz BENSRHAIR[†]

***Abstract*—Feature matching is an important and crucial task in computer vision. Unfortunately, the features extracted from images are usually redundant or irrelevant, leading to many ambiguities and false positives during matching. This work provides an exhaustive study of state-of-the-art feature filtering strategies investigating different criteria, such as the contrast level, inner primary ratio, entropy, saliency and median value, in the purpose of discarding irrelevant features. The evaluation on the well-known Oxford-5k dataset demonstrates the relevance of these criteria to reduce the amount of data involved in the matching step and the number of false positives, thus leading to faster and more accurate matching with almost no loss in performance. Our results are also compared to convolutional neural networks (CNNs), providing extensive experiments that corroborate recently published work.**

***Index Terms*—SIFT, Feature filtering, Pruning, Entropy, Inner primary ratio (IPR), Saliency, Convolutional Neural Networks (CNN).**

I. Introduction

Since the work of Krizhevsky *et al.* in [1] with AlexNet for visual recognition in ILSVRC (ImageNet Large-Scale Visual Recognition Challenge), the current trend in computer vision has been focusing on deep learning, especially convolutional neural networks (CNNs). For these CNN-based methods to perform well, a large amount of data and time are needed to train the network (e.g. 1.2 million training images, 50,000 validation images, and 150,000 testing images with roughly 1000 images in each of 1000 categories for ILSVRC). One solution to these drawbacks is provided by the so-called transfer learning method, where one uses a pre-trained (generic) network and fine-tunes its weights in order to address a specific task at hand [2]. This procedure results in a black box framework for which one does not control the various treatments carried out or its behavior. Also, even if it requires less data than when starting from scratch, a large amount of data is still needed in practice. For all these reasons, researchers have been recently re-investigating methods like Scale Invariant Feature Transform (SIFT), since it overcomes the aforementioned drawbacks. Moreover, recent studies corroborate connections between SIFT and CNN [3], with comparable performances [4].

The SIFT method [5] is still considered a reference in computer vision algorithms due to its robustness to geometric and photometric transformations, as well as the high discriminative power of its 128-element vector descriptors. All those properties make the SIFT method to be preferred to other image matching methods [6]. For computer vision applications such as object recognition [7], image retrieval [8] and many others, feature matching is a very important task. In the naive approach, namely the *exhaustive search*, finding matches between two images consists in comparing each feature of the *first* image to all features from the *second* looking for their nearest neighbor in the descriptor space.

Local descriptors methods such as the SIFT method, often leads to large features databases which, combined with its descriptors dimensions, poses two problems: storage and time consuming of the nearest neighbor search. One idea to speed up the matching process is to reduce descriptors size. This can be done in several ways such as PCA-SIFT [9], Reduced-SIFT [10] or SURF [11]; see for instance [12] for a comparative study. Another way to speed up matching is to look for approximated nearest neighbors (ANN) instead of the real ones. The binarization of descriptors [13], [14] or binary features [15] allow also a phenomenal saving of time during the matching through the use of the Hamming distance, instead of the Euclidean distance usually used to compare real or integer value descriptors; the price to pay is usually a reduced performance compared to the latter. These descriptors also offer a huge gain in storage precisely because binary words take up less space than integer or real ones.

Another major approach, leading to reduction of storage space and matching speed-up is feature filtering. In practice, many of the found matches are false positives [16] and can skew the results when their number is too high, even with the use of a robust geometric fitting method such as RANSAC [17]. Feature filtering is a very simple way of reducing the number of false positives by eliminating irrelevant features according to a certain criterion. Most

[⋆†]S. Konlambigue is with both DATAHERTZ, R&D, Troyes (France), and Normandie Université, University of Rouen Normandy, LITIS Lab, Rouen (France) silvere.konlambigue@datahertz.fr

[⋆]J.-B. Pothin is with DATAHERTZ, R&D, Troyes (France) jean-baptiste.pothin@datahertz.fr

[†]P. Honeine is with Normandie Université, University of Rouen Normandy, LITIS Lab, Rouen (France) paul.honeine@univ-rouen.fr

[†]A. Bensrhair is with Normandie Université, INSA Rouen Normandy, LITIS Lab, Rouen (France) abdelaziz.bensrhair@insa-rouen.fr

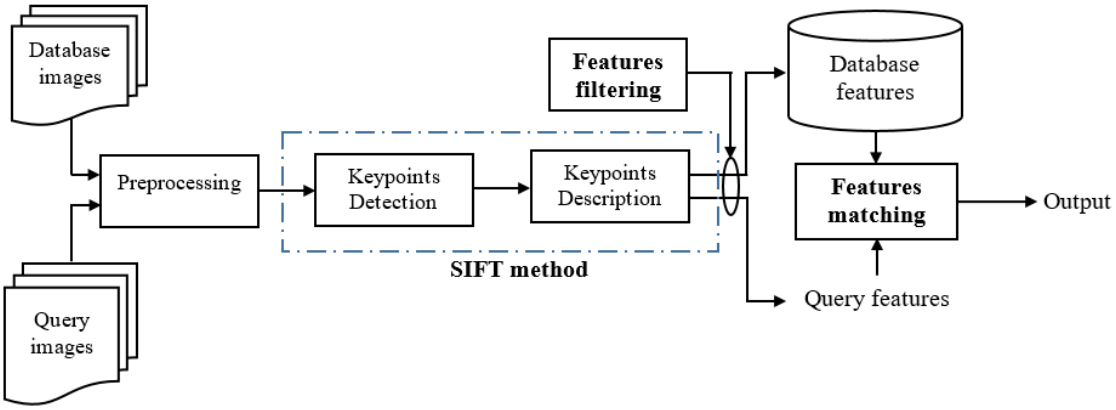

Fig. 1. The SIFT-based image matching schema with the features filtering block as a post-processing step.

prominent methods are based on entropy [18], median [14] value, contrast level [19] and several others. As mentioned, beyond reducing the number of false positives, feature filtering has a double advantage: reducing the number of features involved in the matching process and thus accelerating it and gaining in storage space due to this reduced number of features. These filtering criteria are relatively simple to implement, very inexpensive and can be used upstream to the aforementioned methods (reduced-size descriptors, binary descriptors and ANN) to further speed-up the matching task.

In this paper, we investigate various feature filtering criteria, implement them and evaluate their performance, in order to allow the readers to make a better choice according to their needs and the compromise they are ready to accept. Experiments are conducted on the Oxford-5k dataset [8]. This dataset is quite challenging due to the amount of features and the substantial variations in scale, viewpoint, and lighting conditions. The feature matching strategy used in the experiments is the nearest neighbor distance ratio (NNDR) [5] and performance is evaluated using the mean average precision (mAP) measure [20]. The conducted study provides an in-depth comparative analysis with CNN results and other state-of-the-art methods [4], [21], [22]. This paper is organized as follows: in Section II, we give a brief review of the SIFT method and the matching process. Filtering criteria and their operations are presented in Section III. Section IV presents the dataset, the experimental settings, and the performance evaluations. Section V concludes this paper.

II. SIFT-based Image Matching

The SIFT method is one of the most investigated computer vision algorithms [5]. It operates in two steps. In the first step, called *keypoints detection*, a given image is processed to extract distinctive points using the *Difference-of-Gaussian* (*DoG*) detector. Looking for these points at different scales, by constructing a Gaussian scale-space, allows the SIFT method to be robust to scale changes. Note that the Gaussian scale-space computation is very time-consuming, which can be overcome using for instance extended box filters instead of commonly used Gaussian filters [23]. A keypoint is represented as a tuple $(x, y, \sigma, \theta, \mathcal{R})$ where x and y are its coordinates in the image, σ the scale, θ the orientation and $\mathcal{R}$ the normalized image patch around the keypoint.

In the second step, called *keypoints description*, 16 small 8-bin orientation histograms are computed from 4×4 sub-regions covering the patch $\mathcal{R}$, and then concatenated to form the 128-element keypoint descriptor: $\mathbf{d}=\{d_0, \ldots, d_{127}\}$. Histograms are weighted by the gradient magnitude and a Gaussian window to give less emphasis to points that are far from the keypoint position. The gradient magnitude $m(x, y)$ and orientation $\phi(x, y)$ are computed as follows:

$$m(x,y) = \sqrt{\mathcal{R}_x(x,y)^2 + \mathcal{R}_y(x,y)^2}, \quad (1)$$

$$\phi(x,y) = \arctan\Big(\frac{\mathcal{R}_y(x,y)}{\mathcal{R}_x(x,y)}\Big), \quad (2)$$

with

$$\mathcal{R}_x(x,y) = \mathcal{R}(x+1,y) - \mathcal{R}(x-1,y), \quad (3)$$

$$\mathcal{R}_y(x,y) = \mathcal{R}(x,y+1) - \mathcal{R}(x,y-1). \quad (4)$$

As $\mathcal{R}$ is already normalized by σ and θ, the resulting descriptor is thus scale and rotation invariant.

To find matches between two images, especially in the *exhaustive search* approach, one has to compare each feature from the first image to all the features from the second, looking for their nearest neighbors in descriptor space. The nearest neighbors are typically evaluated using metrics such as the Euclidean distance. In the NNDR strategy, a match between two features is accepted if and only if the ratio between the distance of the first and the second nearest neighbors is lower than some threshold, namely

$$\frac{d_{nn_1}}{d_{nn_2}} < \tau, \quad (5)$$

where $0 < \tau < 1$ denotes a predefined threshold.

III. feature filtering Criteria

In this section, some criteria and their technical details are presented in order to eliminate irrelevant SIFT features. See Fig. 1 for an end-to-end image retrieval system overview.

A. Contrast-based method (CP)

In the SIFT method [5], weak contrast points, points with low DoG values, are filtered-out with respect to a threshold value. This filtering strategy improves the robustness of the detected features. In [19], Foo *et al.* took advantage of this contrast-based filtering strategy and introduced a new pruning scheme. Instead of setting up a threshold value, they suggested to select the top N most significant keypoints ranked by their contrast values. Note that this pruning scheme only affects images that have more than N features.

B. IPR-based method

Treen *et al.* introduced in [24] a new filtering criterion based on the *inner primary ratio* (IPR) value formulated for a 128-element descriptor as:

$$\mathrm{IPR} = \frac{d_{40}^2 + d_{48}^2 + d_{72}^2 + d_{80}^2}{\sum_{i=0}^{127} d_i^2}. \tag{6}$$

High IPR values indicate that the four inner elements (i.e., in the numerator) capitalize most of the descriptor energy, while low IPR values indicate that the energy is distributed on each bin of the descriptor. Only keeping low IPR descriptors reduces ambiguities and false matches from the matching. For example in [24], the IPR threshold value is set to $\mathrm{IPR}_{\mathrm{th}} = 0.235$ after extensive experiments on a synthetic benchmark.

C. Entropy-based method

The entropy is a metric from information theory that quantifies the amount of information contained in a message. In [18], Zivkovic *et al.* evaluated the amount of information hold in a 128-element vector descriptor using the definition:

$$H_1(\mathbf{d}) = -\sum_{i=0}^{127} p_i(\mathbf{d}) \log_2 p_i(\mathbf{d}), \tag{7}$$

with

$$p_i(\mathbf{d}) = \frac{d_i}{\sum_{k=0}^{127} d_k}. \tag{8}$$

On the other hand, Dong *et al.* considered in [25] another approach for estimating the descriptor entropy. In their definition, as the descriptor is integer value $d_k \in \{0, \ldots, 255\}$, the entropy is evaluated in the form:

$$H_2(\mathbf{d}) = -\sum_{i=0}^{255} \tilde{p}_i(\mathbf{d}) \log_2 \tilde{p}_i(\mathbf{d}), \tag{9}$$

with

$$\tilde{p}_i(\mathbf{d}) = \frac{|\{k \mid d_k = i\}|}{128}, \tag{10}$$

where $|\cdot|$ represents the number of elements in the set.

Fig. 2. Example of a saliency map: *left* – original image, *right* – saliency map (overlay), obtained using Itti *et al.* algorithm [26]. The saliency values increase here from blue to red.

D. Saliency-based method

In a field of view, points that stand out from their surround, namely the salient points, visually captivate attention (focus-of-attention) in the scene. Given an input color image, the method proposed by Itti *et al.* in [26] computes a saliency map, namely scalar quantities at each location in the image. This map is computed by combining feature maps from a multi-scale feature extraction method, the center-surround, on intensity, color channels and orientation images derived from the input image; see Fig. 2, obtained using the code provided by Harel *et al.* in [27].

In order to keep only the most relevant features, one has to select features with the highest saliency values. The method was used in [28] on the SURF features in order to filter out irrelevant features by classifying them as *significant* (i.e., highly salient) or *insignificant* (i.e., less salient). The saliency threshold value for this binary classifier varies from 0.2 to 0.3. In [29], [30], the saliency-based feature filtering was used on SIFT features. More specifically, instead of defining a global threshold for all images as in [28], the threshold is made depend on each image and set to 3 times the average salience of the image in [29]. In [30], no threshold was set but features were pruned in order to select the most relevant, such that the ratio between the kept features to their initial number meets the user-defined ratio, varying from 10% to 40%.

E. Filtering by descriptors median value

In the SIFT method [5], features points located along the edges are detected based on their principal curvatures and eliminated because they are unstable and prone to false matches. Despite this strategy, Zhou *et al.* noted

TABLE I

THE mAP (%) PERFORMANCE EVALUATION ON THE OXFORD-5K DATASET. THE VALUES FOR EVERY LANDMARK (L1 TO L11) ARE GIVEN WHEN AVAILABLE, AS WELL AS THEIR AVERAGE (COLUMN mAP) AND THE REDUCTION LEVEL ρ (%) OVER THE DATABASE. THE FIRST FIVE LINES ARE TAKEN FROM THE RESPECTIVE REFERENCES. BOVW[X] = BAG OF VISUAL WORDS [DICTIONARY SIZE]. SP = SPATIAL VERIFICATION. X = MULTIPLE ASSIGNMENT + WEIGHTING HAMMING EMBEDDING + SP.

Methods	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	mAP	ρ
SIFT+BoVW[1M]+SP [21]	-	-	-	-	-	-	-	-	-	-	-	64.50	-
SIFT+BoVW[500k]+td-idf [32]	-	-	-	-	-	-	-	-	-	-	-	61.30	-
SIFT+BoVW[20k]+X [33]	-	-	-	-	-	-	-	-	-	-	-	68.50	-
SIFT+BoVW[4096] [4]	36.56	34.25	36.35	41.43	38.17	40.51	69.80	56.03	25.15	56.58	60.24	45.01	-
SIFT [4]	44.90	51.87	51.12	67.24	65.88	67.04	69.24	87.44	27.58	99.18	59.11	62.79	-
CNN [4]	40.36	40.74	39.64	36.18	47.51	45.30	80.65	58.35	27.90	76.45	88.13	52.84	-
Our SIFT baseline (NNDR)	58.64	53.75	50.75	66.59	66.42	61.31	73.18	100	19.75	100	74.68	65.91	0
SIFT+CP[N=5000]	56.33	53.33	50.44	68.42	66.73	61.56	73.21	99.37	19.92	100	74.18	65.77	10.44
SIFT+IPR[th=0.25]	55.88	53.14	52.04	67.32	66.39	58.40	71.95	100	18.74	100	71.42	65.03	10.57
SIFT+Entropy H1[th=5.4]	57.11	51.28	50.43	70.04	66	58.13	69.92	99.37	19.13	100	71.33	64.79	10.07
SIFT+Entropy H2[th=3.5]	58.53	52.60	49.54	69.42	65.70	57.93	69.72	98.04	18.63	100	71.18	64.66	11.45
SIFT+Saliency[th=0.15]	55.79	52.92	49.94	68.92	64.04	62.27	73.69	96.26	19.40	100	70.67	64.90	11.75
SIFT+Median[th=0.007]	58.84	53.33	48.61	68.58	65.61	58.70	70.24	97.84	18.59	100	71.20	64.69	10.52

in [14] that some of these points are still present and degrade the matching accuracy. Similar to Treen *et al.* [24], Zhou *et al.* also noted that, for these keypoints, most of the coefficients of their descriptors bins are of low magnitude leading to low median values. One is then able to reduce irrelevant points by filtering out low median value features. In [14], the threshold value is set to $\mathrm{M_{th}} = 5.5$ for integer values descriptors in the range [0 , 512].

IV. EXPERIMENTS AND RESULTS

A. Dataset

To effectively evaluate performance, experiments must be conducted on a large and quite challenging dataset. For this purpose, we evaluate each of the filtering criteria on the Oxford-5k dataset [8] implementing an image retrieval system. This dataset contains 5062 high resolution images distributed in 11 different landmarks, denoted L1 to L11, and indexed by 55 query images, 5 images per landmark with associated region of interest (ROI). Each query has four ground-truth labels:

- *Good* – the ROI is clearly visible
- *Ok* – more than 25% of the ROI is visible
- *Junk* – less than 25% of the ROI is visible (including high occlusion and high distortion)
- *Absent* – the ROI is not present (namely, images that are not in the first three classes)

In performance assessment, according to the protocol of the Oxford-5k dataset, *Good* and *Ok* labels are considered true positives, while *Absent* labels are false positives. *Junk* labels are considered neutral and do not affect the performances. Considering the SIFT method, these 5062 database images yield about 17 millions features. Note that our experiments implement the Multiple Match Removal (MMR) method [31].

B. Performance measure

To evaluate the retrieval performance on the Oxford 5k dataset, the mean average precision (mAP) is used, as described in the following.

In the matching stage, each query is compared to all of the database images. A descending ranked list is then returned according to the level of similarity defined as:

$$sim(I_q, I_s) = \frac{m_{q,s}}{m_q}, \tag{11}$$

where $m_{q,s}$ is the number of features that satisfy (5), between query image I_q and database image I_s, and m_q is the number of the query features. Note that $sim(\cdot,\cdot) \in [0,1]$ with $sim(\cdot,\cdot) \approx 1$ (resp. ≈ 0) means images are very similar (resp. dissimilar).

From the ranked list of images, one evaluates the average precision (AP) for query q as follow [20]:

$$\mathrm{AP}(q) = \frac{\sum_{r=1}^{N} P(r) \times rel(r)}{\text{number of relevant images}}, \tag{12}$$

with r the retrieved image rank, N the number of retrieved images, $rel(r)$ an indicator equaling 1 if the image at rank r is relevant relatively to q and 0 otherwise. $P(r)$ is the precision at cut-off r defined as the ratio of the number of relevant retrieved images at rank r, over r.

As the AP is calculated for each query q, the mAP is obtained by averaging over all queries:

$$\mathrm{mAP} = \frac{\sum_{q=1}^{N_q} \mathrm{AP}(q)}{N_q} \times 100\%, \tag{13}$$

where N_q is the total number of query images.

To compare the methods we evaluate the change in mAP, relatively to the baseline mAP, against the percent-

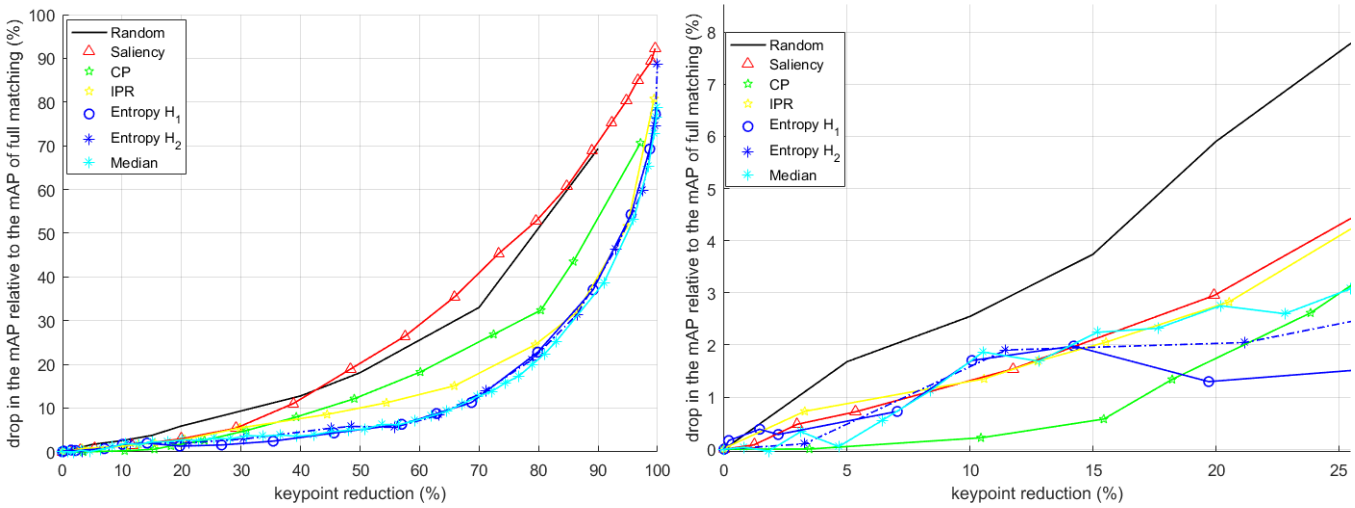

Fig. 3. Performance comparison of filtering criteria. The right figure is a zoom in for $\rho \leq 25\%$.

age of filtered data. The reduction level ρ over storage in the database is defined as:

$$\rho = \left(1 - \sum_s \frac{n_s}{N_s}\right) \times 100\%, \tag{14}$$

where n_s, resp. N_s, is the number of remaining features, resp. the total number of features before the filtering, from the database image I_s.

C. Performance evaluation

The results of the experiments are given in TABLE I. In the experiments, the baseline is the original SIFT. Its mAP value on the Oxford-5k dataset is mAP = 65.91%, which is underlined in TABLE I. In this table, the values in columns L1-L11 represent the mAP values according to each landmark, and the columns "mAP" and ρ are respectively the averages over all landmarks and the reduction rates defined in (14). The first six lines recall the state of the art; see [3] for a complete comparison of SIFT and CNN based methods. Note that we match and even improve these performances without using any codebook.

In the literature, the reduction rates often put forward are around 12%. Fig. 3 shows the loss in mAP according to the percentage of reduction for each of the aforementioned methods. The right figure is a simple zoom out of the left one. From this figure, three observations can be made. First, none of the filtering criteria improves the mAP. Each of them results in a loss of performance depending on the rate of reduction. Secondly, for reduction rates of less than about 15%, most methods seem equivalent to each other, except the contrast-based method (CP) which appears to be the best one. Finally, at higher reduction rates, the entropy and median provide the lowest drop in mAP. This is expected because it is well understood that a weak entropy descriptor must have a low median. The IPR-based method is less good but joins those of the entropy when reduction factor is important. Note that for reduction rates greater than 40%, the saliency-based approach is worst than random filtering.

In TABLE I we give details of mAP for thresholds that do not exceed 12% threshold. Also, in our experiments, an IPR threshold value $\text{IPR}_{\text{th}} = 0.235$ allows a data reduction of about 12% with only 1% loss of mAP and therefore corroborates the recommendations made in [24]. It is worth noting the difference in thresholds for both types of entropy. Although it is the same descriptor entropy that is evaluated in both approaches, the difference in the thresholds values comes from the nature (integer or real) of the descriptor, as described in (7) and (9) in Section III-C. Through experiments, we have observed that one should pay attention to this detail as it can have a highly negative impact on the performance; see Fig. 4.

V. Conclusion

In this paper we have reviewed a number of criteria proposed in the state of the art that can be used for SIFT feature filtering. The obtained results match and even improve these performances of state-of-the-art methods, including deep neural networks with CNN. Although some criteria seem equivalent in terms of mAP performance, according to a certain rate of reduction, two important aspects need to be taken into account when choosing a criterion: the ease of implementation and the cost in computing time. On these bases, we advise the criteria based on the entropy, for the following reasons: it is directly computed from the descriptor, practical when one uses an already implemented version of a computer vision algorithm, and also because a filtering according to this criterion is based on the richness of the descriptor.

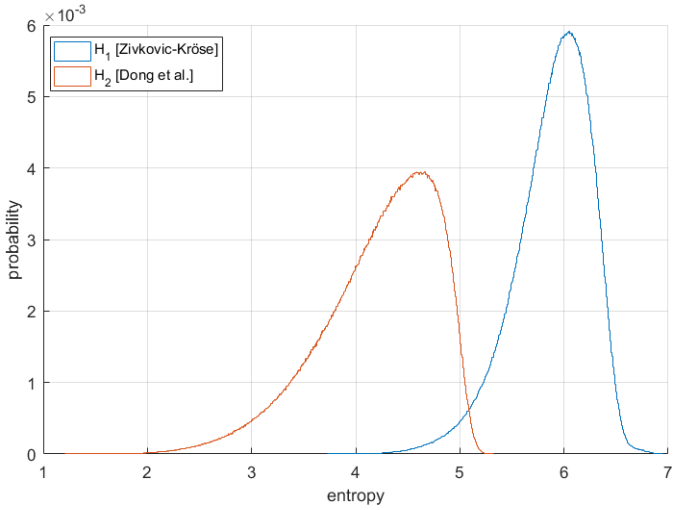

Fig. 4. Histogram distribution of entropy H_1 and H_2 extracted from the Oxford-5k dataset.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] M. Liu, R. Chen, D. Li, Y. Chen, G. Guo, Z. Cao, and Y. Pan, "Scene recognition for indoor localization using a multi-sensor fusion approach," *Sensors*, vol. 17, no. 12, p. 2847, 2017.

[3] L. Zheng, Y. Yang, and Q. Tian, "SIFT meets CNN: A decade survey of instance retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1224–1244, 2017.

[4] V. D. Sachdeva, J. Baber, M. Bakhtyar, I. Ullah, W. Noor, and A. Basit, "Performance evaluation of SIFT and Convolutional Neural Network for image retrieval," *Performance Evaluation*, vol. 8, no. 12, 2017.

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

[6] H. Chatoux, F. Lecellier, and C. Fernandez-Maloigne, "Comparative study of descriptors with dense key points," in *23rd International Conference on Pattern Recognition (ICPR)*, 2016, pp. 1988–1993.

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

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[9] Y. Ke, R. Sukthankar *et al.*, "PCA-SIFT: A more distinctive representation for local image descriptors," *CVPR (2)*, vol. 4, pp. 506–513, 2004.

[10] L. Ledwich and S. Williams, "Reduced SIFT features for image retrieval and indoor localisation," in *Australian conference on robotics and automation*, vol. 322, 2004, p. 3.

[11] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded Up Robust Features," in *European conference on computer vision*, 2006, pp. 404–417.

[12] J. Wu, Z. Cui, V. S. Sheng, P. Zhao, D. Su, and S. Gong, "A comparative study of SIFT and its variants," *Measurement science review*, vol. 13, no. 3, pp. 122–131, 2013.

[13] K. A. Peker, "Binary SIFT: Fast image retrieval using binary quantized sift features," in *9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2011, pp. 217–222.

[14] W. Zhou, H. Li, R. Hong, Y. Lu, and Q. Tian, "BSIFT: toward data-independent codebook for large scale image search." *IEEE Transactions on Image Processing*, vol. 24, no. 3, pp. 967–979, 2015.

[15] J. Heinly, E. Dunn, and J.-M. Frahm, "Comparative Evaluation of Binary Features," in *European Conference on Computer Vision (ECCV)*, 2012.

[16] W. Hartmann, M. Havlena, and K. Schindler, "Predicting matchability," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 9–16.

[17] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[18] Z. Zivkovic and B. Kröse, "On matching interest regions using local descriptors - can an information theoretic approach help?" in *Proceedings of the British Machine Vision Conference (BMVC)*, 2005, pp. 50–58.

[19] J. J. Foo and R. Sinha, "Pruning SIFT for scalable near-duplicate image matching," in *Proceedings of the 18th conference on Australasian database*, 2007, pp. 63–71.

[20] S. Nikolopoulos, S. Zafeiriou, I. Patras, and I. Kompatsiaris, "High order pLSA for indexing tagged images," *Signal Processing*, vol. 93, no. 8, pp. 2212–2228, 2013.

[21] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[22] H. Jégou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search - extended version," Research Report 6709, 2008.

[23] S. Konlambigue, J.-B. Pothin, P. Honeine, and A. Bensrhairt, "Fast and accurate gaussian pyramid construction by extended box filtering," in *26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 400–404.

[24] G. Treen and A. Whitehead, "Efficient SIFT matching from keypoint descriptor properties," in *Workshop on Applications of Computer Vision (WACV), 2009.* IEEE, 2009, pp. 1–7.

[25] W. Dong, Z. Wang, M. Charikar, and K. Li, "High-confidence near-duplicate image detection," in *Proc. of the 2nd International Conference on Multimedia Retrieval*, 2012, pp. 1–8.

[26] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[27] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2007, pp. 545–552.

[28] H. M. Sergieh, E. Egyed-Zsigmond, M. Doller, D. Coquil, J.-M. Pinon, and H. Kosch, "Improving SURF image matching using supervised learning," in *8th International Conference on Signal Image Technology and Internet based Systems (SITIS)*. IEEE, 2012, pp. 230–237.

[29] D. Zhao, J. Wang, J. Wan, and T. Xiao, "Fast SIFT scene matching algorithm based on saliency detection and frequency segmentation for downward-viewing images," in *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering*, 2013.

[30] D. Awad, V. Courboulay, and A. Revel, "Saliency filtering of SIFT detectors: Application to cbir," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2012, pp. 290–300.

[31] K. Mikolajczyk and C. Schmid, "Indexing based on scale invariant interest points," 2001, pp. 525–53.

[32] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Descriptor learning for efficient retrieval," in *European Conference on Computer Vision*, 2010, pp. 677–691.

[33] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1169–1176.