Structure Oriented Local Image Feature Description

Wangsheng Yu1,*, Xianxiang Qin1, Peng Wang1, Zhiqiang Hou2 and Sugang Ma2
1Information and Navigation College Air Force Engineering University, Xi’an, China
2Xian University of Posts & Telecommunications, Xi’an, China

*Corresponding author

Abstract. In this paper, we proposed a structure oriented descriptor (SOD) for local image feature description. Different from the traditional methods, we explored the structure information via hierarchical strategy and structure coding elements. Firstly, the support region is partitioned into hierarchical sub-regions according to the intensity orders. Secondly, a pre-designed structure coding image is explored to pooling the features according to the hierarchical sub-regions. The final descriptor is a conjunction of the feature of each sub-region.

We evaluated the proposed descriptor on the public dataset and compared it with the state-of-the-art works. The experimental results indicate that the proposed descriptor is robust to image appearance change and outperform the state-of-the-art works in most cases.

1. Introduction

Image feature descriptor is the mapping result from raw pixel information to image feature space, which is of great importance for image retrieval, object recognition, image mosaic and so on. During the past decades, many image feature descriptors are proposed to achieve discriminative performance while maintaining the robustness to some typical image changes, such as scale, rotation, viewpoint, illumination, and image blur. Among these descriptors, SIFT1 and SURF2 are two influential works. SIFT divides the support region into sub-regions in order to encode more structure information, which improves the discrimination of the feature. It achieves the robustness on scale and rotation via the key-points detection and reference orientation estimation. SURF is an amelioration of SIFT, especially in the efficiency. It owns a comparable performance to SIFT and has a widely application for the acceleration. DAISY3 is essentially a member of SIFT family, which accumulates the gradient orientation histogram according to the partitioned sub-regions. It designs the feature pooling via a Gaussian convolution which increases the efficiency especially for dense feature extraction. More similar works are GLOH4, ASIFT5, CSIFT6 and PCA-SIFT7.

In order to relieve the storage burden and improve the matching efficiency, many binary features are proposed. BRIEF8 is a representation which is known as its high efficiency. It constructs the feature by concatenating the results of binary tests of intensities for several random point pairs in image patch. However, it is sensitive to image rotation and scale change. Some similar works are FREAK9 and ORSI10. CS-LBP11 encodes intensity ordinal information based on local binary pattern (LBP)12. The reported performance is superior than SIFT, moreover, it is computationally simpler than the SIFT. Fan et al. proposed a novel RFD13 based on the responses of a set of receptive fields, which are selected from a large number of candidates according to their distinctiveness and correlations in a greedy way.

Recent years, the great success of deep learning in computer vision arouses the learning-based feature description. It can learn a general representation of the image content via a deep network (such as AutoEncoder14, and CNN15). Such features have been widely applied in object recognition16, object
classification\textsuperscript{17}, visual tracking\textsuperscript{18}, and so on. Compared with the hand-crafted descriptor, these learning-based features are more intrinsic representation of image. However, it is time-consuming to pre-train a deep network.

In this paper, we proposed a novel feature descriptor by exploring the structure information via hierarchical strategy and structure coding elements. The experimental results on public dataset validate its superiority.

The whole paper is organized as follows. Section 1 gives a brief introduction the background and inspiration of the proposed work. Then, the main idea and concrete processing course of the proposed work are discussed in Section 2. Section 3 compared the proposed work with some relative descriptors. We conclude the whole paper in Section 4.

2. Proposed Method

In general, a good local feature descriptor should achieve discriminative performance and be robust to appearance changes, such as scale, rotation, viewpoint, illumination, and image blur. So, there are two aspects, the discrimination and robustness, which should be paid much attention to when design image feature descriptors. Many recent works explore the spatial information by dividing the support region into sub-regions to achieve the robustness, which have been proved to be effective. SIFT\textsuperscript{1} divides the support region as regular grids. It improves the robustness to rotation by selecting a reference orientation. MROGH\textsuperscript{19} obtains the sub-regions according to the intensity orders, which are inherent robust to image rotation. In this paper, we combined these two methods and proposed a novel way to pool the feature.

2.1. Affine Invariant Region Detection

During the past decades, researchers proposed a lot of interesting region detectors, such as SIFT, Harris-Laplace, Harris-Affine, Hessian-Laplace, Hessian-Affine, and so on. These detectors are hand-crafted in order to invariant to scale change, rotation and affine variation. The repeatability is a very important criterion to evaluate the performance of feature detectors. Ives \textit{et al.}\textsuperscript{20} compared 12 feature detectors proposed in recent years based on Oxford dataset\textsuperscript{21} according to a new repeatability criterion. The results indicate that when the redundancy is not taken into account the method producing the most detections and the highest repeatability is the Hessian-Laplace, while when considering the non-redundant variant it is SIFT.

However, affine detectors (such as Harris-Affine and Hessian-Affine) can detect elliptical regions that are affine invariant up to a rotation transformation. Besides, these detectors are more efficient than SIFT. So, in this paper, we exploit the Hessian-Affine to obtain the affine invariant regions. A post-processing is introduced to normalize the detected region to a canonical region. We refer to work of Fan \textit{et al.}\textsuperscript{19} for more details about the normalization.

Finally, we obtain the normalized affine invariant support regions with size of 41 × 41 pixels.

![Figure 1. The structure coding image.](image)

2.2. Support Region Partition

For most of the local feature descriptors, a common way to encode the structure information is the support region partition. Both SIFT and HOG\textsuperscript{22} divide the support region into regular rectangular sub-
regions, while GLOH prefers the annular as the support region partition strategy. These issues are employed to increase the discrimination of the descriptors.

In this paper, we partition the support region according to the intensity orders of the pixels. Denote by $R = \{x_1, x_2, ..., x_n\}$ a support region with $n$ pixels, and $I(x_i)$ the intensity of the pixel $x_i$. The $n$ pixels are partitioned into $k$ groups as:

$$R_j = \{x_i \in R: I_{j-1} < I(x_i) < I_j\}, j = 1, 2, ..., k$$

(1)

where $I_{j-1}$ and $I_j$ are the boundaries the $j^{th}$ interval, which is an equally divided part of the intensity range of the support region.

The partition based on intensity orders makes the sub-regions invariant to image rotation. A similar partition method is used in Fan et al.19 The difference is that Fan et al.19 equally divides the support region according to the number of pixels, while in this paper, we equally divide the support region according to the intensity range. The experiment results indicate that our method owns a better discrimination.

Figure 2. Demonstration of constructing the proposed feature descriptor.

2.3. Structure Coding Image Design

Through the support region partition, we obtain the rotation invariant sub-region and encode partial spatial information. In order to make full use of spatial and structure information to improve the
discrimination, we design a structure coding image to further encode the structure information. We refer to Fig.1 for more details about the design of structure coding image.

In Fig.1, we display two types of structure coding image. The left one encodes the support region with 9 elements and the right one with 17 elements. To be mentioned is that the structure coding image is with the same size of support region. Generally speaking, less structure elements will obtain a better robustness, while decrease the discrimination. We will compare these two structure coding images during the evaluation section.

2.4. Feature Pooling

For most of the descriptors (such as SIFT), they calculate the feature of each partitioned sub-region and integrate the results into the final feature. In this paper, we propose a novel feature pooling strategy. We treat the structure coding elements as feature elements, and calculate the statistical histogram of these elements according to the partitioned sub-regions. Such an operation improves the discrimination while maintaining the robustness.

Before the feature pooling, we calculate the reference orientation according to the orientation of each sub region. We exploit the covariance matrix of the normalized coordinates (the origin is the center of the support region) of the pixels to calculate the sub-region orientation. Actually, each orientation has a different weight to the orientation of the support region. So, if there exist a predominant orientation, we take it as the reference orientation; otherwise, we treat the sub-region orientations as vectors and obtain the reference orientation through vector synthesis operation. After the reference orientation obtained, we rotate the structure coding image to the reference orientation, and finish the following process. The final feature descriptor is a conjunction of the feature of each sub-region. We refer to Fig.2 for a demonstration of the whole course of feature descriptor construction.

3. Results

In order to validate the superiority of the proposed feature descriptor, we compare it with some state-of-the-art works via a set of experiments based on the Oxford dataset. The referenced descriptors consist of the float features, such as SIFT and MROGH, and binary features, such as FREAK and ORSI. SIFT and FREAK are two outstanding features while MROGH and ORSI are two similar works to the proposed feature.

Before the evaluation, we first explore a set of experiments for the proposed feature descriptor to obtain the optimal parameters. There are two parameters of the proposed descriptor, the number of the partitioned sub-regions according to intensity orders and the number of the structure coding elements. The former can be assigned different values, such as 4, 8, 16, 32, and so on. The latter can be 9 and 17 according to the demonstration in Fig.1. So there are 8 combinations of the parameters for the proposed descriptor. We evaluate these 8 combinations based on the dataset. The comprehensive experiments indicate that the combination when partition number equals 8 and structure coding elements is 17 obtains a relative better performance. Although when the partition number increases to 16 may improve the performance slightly, we still choose the 8 and 17 as the optimal parameters comprehensively considering the effect and efficiency. We mark the proposed structure oriented descriptor with SOD-136 (the number of partitions times the number of structure coding elements).

Based on the Hessian-Affine detector, we compare the description performance of the descriptors. We adopt the recall versus 1-precision curves as the evaluation criterion in this paper, which is also used in literatures [4], [10] and [19]. We refer to these literatures for the details about the definitions of recall and 1-precision.

Partial experiment results are displayed in Fig.3. The image appearance changes contain viewpoint changes, zoom, rotation, image blur, and illumination changes. The curves indicate that MROGH-192 outperforms than SIFT-128 as float feature descriptors, and ORSI-32 is superior than FREAK-16 as binary feature descriptors. The proposed descriptor, SOD-136, obtains a relative better performance in most of the cases except for the image blur.
We attribute the success of the proposed descriptor to the structure coding image and support region partition based on intensity orders. This designation increases the feature discrimination while maintaining the robustness to the image appearance variation.

![Figure 3](image)

**Figure 3.** Experimental results under various image transformations in the Oxford data set for Hessian-Affine detector.

It should be pointed out that the proposed SOD-136 is slightly efficient than SIFT-128, MROGH-192 and ORSI-32, while less efficient than FREAK-16.

4. Conclusion

In this paper, we proposed a simple but very effective feature descriptor. Different from the existing descriptors, we explored the structure information via hierarchical strategy and structure coding elements. Firstly, the support region is partitioned into hierarchical sub-regions according to the intensity orders. Secondly, a pre-designed structure coding image is explored to pooling the features according to the hierarchical sub-regions. These two issues increase the feature discrimination while maintaining the robustness to the image appearance variation. This feature descriptor can also be used in region-based feature description. We evaluate the effectiveness of the proposed descriptor for region-based feature description under the visual tracking experiments.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No.61703423, 41601436, and 61773396).

References

[1] D. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 2, no. 60, pp. 91-110, 2004.

[2] H. Bay, T. Tuytelaars, and L. V. Gool., "Surf: Speeded up robust features," in Proceedings ECCV. 2006, pp. 404-417.

[3] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, vol. 32, no. 5, pp. 815-830.

[4] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, vol. 27, no. 10, pp. 1615-1630.

[5] J.M. Morel and G. Yu, "ASIFT: a New framework for fully affine invariant image comparison," SIAM Journal on Imaging Sciences, 2009, vol. 2, pp. 438-469.

[6] A.E. Abdel-Hakim, A.A. Farag, "CSIFT: a SIFT descriptor with color invariant characteristics," in Proceedings CVPR. 2006, pp. 1978-1983.
[6] S. Zickler, A. Efros, "Detection of multiple deformable objects using PCA-SIFT," in Proceedings AAAI conference on artificial intelligence. 2007, pp. 1127-1132.

[7] M. Calonder, V. Lepetit, C. Strecha, et al., "BRIEF: binary robust independent elementary features," in Proceedings ECCV. 2007, pp. 778-792.

[8] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in Proceedings CVPR. 2012, pp. 510-517.

[9] X. Xu, L. Tian, J. Feng, et al., "OSRI: a rotationally invariant binary descriptor," IEEE Transactions on Image Processing. 2014, vol. 23, no. 7, pp. 2983-2995.

[10] M. Heikkila, M. Pietikainen, and C. Schmid, "Description of interest regions with local binary patterns," Pattern Recognition, 2009, vol. 42, no. 3, pp. 425-436.

[11] T. Ojala, M. Pietikainen, and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," Pattern Recognition, 1996, vol. 29, no. 1, pp. 51-59.

[12] B. Fan, Q. Kong, T. Trzcinski, et al., "Receptive fields selection for binary feature description," IEEE Transactions on Image Processing. 2014, vol. 23, no. 6, pp. 2583-2595.

[13] N.Y. Wang and D. Yeung, "Learning a deep compact image representation for visual tracking," in Proceedings NIPS. 2013, pp. 809-817.

[14] L. Zhang and P.N. Suganthan, "Visual tracking with convolutional random vector functional link network," IEEE Transactions on Cybernetics. 2017, vol. 47, no. 10, pp. 3243-3253.

[15] J.C. Piao, H.S. Jung, C.P. Hong, et al., "Improving performance on object recognition for real-time on mobile devices," Multimedia Tools and Applications, 2016, vol. 75, no. 16, pp. 9623-9640.

[16] A. Romero, C. Gatta, G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," IEEE Transactions on Geoscience and Remote Sensing, 2016, vol. 54, no. 3, pp. 1394-1362.

[17] L. Wang, W. Ouyang, X. Wang, "Visual tracking with fully convolutional networks," in Proceedings ICCV. 2015, pp. 3119-3127.

[18] B. Fan, F. Wu, and Z. Hu, "Rotationally invariant descriptors using intensity order pooling," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, vol. 34, no. 10, pp. 2031-2045.

[19] R.O. Ives, D. Mauricio, M. Jean-Michel, "Comparing feature detectors: A bias in the repeatability criteria," in Proceedings ICIP. 2015, pp. 3024-3028.

[20] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in Proceedings CVPR. 2005, pp. 20-25.

[21] K. Mikolajczyk, T. Tuytelaars, C. Schmid, et al., "A comparison of affine region detectors," International Journal of Computer Vision, 2005, vol. 65, no. 1-2, pp. 43-72.