A SEMI-HANDCRAFTED KEYPOINT DETECTOR WITH DISCRIMINATIVE FEATURE ENCODING

Yurui Xie[⋆†], *Ling Guan*[⋆]

[⋆]Ryerson Multimedia Research Laboratory, Toronto, Canada
[†]Chengdu University of Information Technology, Chengdu, China

ABSTRACT

Most previous handcrafted keypoint methods focus on designing specific structural patterns using human-defined knowledge. These methods, however, ignore the fact that whether they have enough flexibility to harvest diverse local structures. Recently, the semi-handcrafted approaches based on sparse coding have emerged as a new trend of alleviating the above issue. And yet, the intrinsic relationships of keypoints have not been explored actively, which may lead to the ambiguity of feature codes for further analysis. To tackle this problem, in this paper, we introduce a novel semi-handcrafted keypoint detector through a scheme of discriminative feature representations (SDFR). Specifically, we cast keypoint detection as an optimization problem on a visual dictionary that explicitly models the visual relationships of feature points to preserve the consistency of similar features and distance dissimilar ones. Further, we propose an iterative solver for the SDFR model. Experimental results on challenge benchmarks demonstrate that the proposed method performs favorably against state-of-the-art in literature.

Index Terms— Keypoint, feature representation, sparse coding, visual dictionary

1. INTRODUCTION

The extraction of keypoints is a fundamental component in various applications of computer vision, such as image matching, pose estimation, automatic driving and more. For years, among existing works, most researches focus on designing specific structural patterns using manually defined knowledge. Despite inspiring progress over the past decades, it remains unclear if such handcrafted detectors can offer enough flexibility to accommodate diverse structural information implied in images. In other words, the roots behind handcrafted detectors usually stem from our intuitions, it is hard to cover all the potential local structures only relying on the limited support of our prior knowledge.

This work is supported by the National Natural Science Foundation of China (No. 61806028), and the Fund of China Scholarship Council (No. 201908510028). The authors gratefully acknowledge the infrastructure support from Ryerson University.

In recent years, the advances of convolutional neural networks (CNNs) motivate the possibility of performing keypoint extraction via a data-driven manner. Essentially, the attractive merit of these learning based methods is that it provides a natural mechanism for capturing various local structures. Although CNNs based approaches have gained impressive performances, state-of-the-art learning based detectors [1, 2, 3] still use the outputs of off-the-shelf handcrafted detectors as the starting point in training to avoid using raw image pixels, which are likely of poor quality, as the input. That is to say, handcrafted detectors could provide higher confidence positive samples to let the models know what keypoints look like. Nevertheless, the drawback of these CNNs based detectors is that they are inevitably restricted to specific structural patterns, where the visual information of extracted keypoints is similar to the training samples.

In order to adequately localize diverse keypoints, [4] is a recent pioneering work towards a solution. This method utilizes sparse coding to capture various keypoints in a semi-handcrafted manner. However, since each keypoint is independently detected, intrinsic correlations of feature points have not been explored, which may weaken the discriminative capability of feature codes for keypoint recognition tasks. Differently, in this paper, we formulate keypoint detection as a discriminative feature encoding scheme that indeed strengthens the visual relationships of keypoints to favorably explore distinctive and diverse local structures.

To summarize, the major contributions of this work are twofold. (a) We propose a novel formulation of discriminative feature encoding that consolidates the visual relationships of feature points for exploiting distinctive and diverse keypoints. (b) We decompose the objective function of SDFR into multiple sub-problems and further propose an iterative solver. Experiments on benchmarks suggest that SDFR achieves favorable performances compared with state-of-the-art methods.

2. RELATED WORK

Existing researches on keypoint detection can be generally grouped into three categories: handcrafted, learning based and semi-handcrafted methods. In the first category, manually defined structures [5, 6, 7, 8, 9, 10] are typically leveraged

to capture specific visual patterns. For example, early works [5, 6] localize corners and blobs through computing image derivatives. Then SIFT [7] and SURF [8] look for blobs by designing specific intensity patterns. KAZE [11] and its enhanced version AKAZE [12] use a nonlinear scale space to keep the boundaries during detection. Also, there are previous efforts [13, 14, 10] that model more complicated blob-like structures, such as local symmetries [13], edge orientations [14], spiral structures [10], to extract various feature points.

With the fast pace of deep learning, attention has been paid to the idea of learning strategy. TILDE [1] is an earlier attempt that uses CNN architecture to identify keypoints. However, it requires a large amount of well-aligned SIFT [7] features from the same scenes. Later, [15] develops a covariant feature detector that treats keypoint detection as a regression problem. [3] introduces a shallow network to extract distinctive regions using the outputs of TILDE [1]. Besides, [2] employs handcrafted features as anchors and combines them with CNNs to enhance the repeatability of keypoint detection. Despite the competitive performance that has been exhibited, these methods are still built upon handcrafted detectors.

More recently, semi-handcrafted methods [4, 16, 17] are emerging as an important tool to explore diverse local structures. Specifically, SCK [4] is the first work that utilizes the standard sparse coding to describe various visual patterns. Thanks to the reconstruction property of sparse coding, the subsequent work [16] is further extended to handle the scale and rotation changes. Moreover, SAKD [17] incorporates a group-sparsity regularization to mine sharable structural information from the same scenes. To this end, the SDFR detector belongs to the family of semi-handcrafted methods, and beyond that SDFR explicitly models the visual relationships among keypoints for steadily strengthening the discriminative power of feature encoding for keypoint recognition.

3. SEMI-HANDCRAFTED KEYPOINT DETECTOR WITH DISCRIMINATIVE FEATURE ENCODING

Sparse coding [18] has the capability to determine feature representations of input images using a pre-defined dictionary. Sparsity allows robustness to distractors and its reconstruction nature provides a way towards describing various visual patterns. Inspired by the two valuable attributes, we propose a scheme of discriminative feature encoding to explore high-quality keypoints as follows.

3.1. Discriminative Feature Encoding for Image Patches

Given an input image, a Gaussian filter is first utilized to remove the noise. Then the image is partitioned into $n \times n$ patches, and each patch is reshaped to a n^2-dimensional normalized vector. To overcome the weakness of handcrafted detectors which are restricted to specific structural patterns, the key insight to the proposed approach is that the detector is expected to discover diverse and distinctive local structures. Therefore, a visual dictionary is naturally suited for this purpose. Specifically, we propose a discriminative feature encoding formulation to cast keypoint detection as an optimization problem via a semi-handcrafted manner:

$$\min_{C} \parallel F - D \cdot C \parallel_F^2 + \alpha \sum_{i=1}^{N} \sum_{j \neq i}^{N} \parallel c_i - c_j \parallel_F^2 \cdot S_{ij} + \beta \parallel C \parallel_F^2 \quad (1)$$

where F denotes the vectorization of patches, D is a visual dictionary. Based on the Fourier series theory [19], we employ sampled sine functions to form a part of D for faithfully synthesizing an arbitrary signal by summation approximation. Additionally, a set of corners [20] generated with 2D-DCT bases multiplied by a window-function are further concatenated to the dictionary, thereby tightening the distribution of eigenvalues of D for improving the discriminative power of feature codes. $C = [c_1, c_2, ..., c_N]$ denotes the feature codes for all the patches, and N is the total number of patches in an image. S_{ij} is the pairwise similarities in the visual space. α and β are the trade-off parameters.

The first term $\parallel F - D \cdot C \parallel_F^2$ is the reconstruction residual that bridges the gap between dictionary atoms and image patches. Therefore, diverse structural patterns are easily highlighted through strong responses that appear on dictionary atoms. For keypoint recognition, we suggest that the visual dictionary should not only reconstruct patches but also produce informative feature codes for better distinguishing various local structures. Diving deeper into this, we argue that the mutual relationship of feature points is important to exploiting distinctive keypoints. Concretely, the feature codes should reflect the correlations of similar keypoints by sharing dictionary atoms as well as distance the dissimilar ones. To achieve this goal, we explicitly incorporate the second term of Eq.(1) as an affinity constraint to generate more discriminative feature codes, S_{ij} denotes the pairwise similarities calculated by:

$$S_{ij} = \exp\left[-\frac{dist(f_i, f_j)}{\sigma}\right], \quad i, j = 1, 2, ..., N \quad (2)$$

where $dist(f_i, f_j)$ measures the Euclidean distance between the i-th and j-th patch vectors. σ is used to adjust the weight decay ratio for scoring similarities. Instead of representing each patch separately, this term drives the detector to encode all the patches in an image jointly. More importantly, it enables the proposed model to characterize similarity relationships of feature codes transferred from the visual space. Note that a similar idea has been applied to the image classification task [21]. However, unlike the prior work, the SDFR model not only incorporates a different sparsity as a regularization to further alleviate the ambiguity of feature codes, but also formulates a new objective function that is differentiable and

computationally feasible for encoding dense patches in keypoint detection scenarios.

Feature codes reveal the meaningful responses of patches on dictionary atoms, where L_1-norm is typically employed to measure the sparsity of feature codes. However, L_1-norm is computationally expensive and sensitive to the variance of inputs [21, 22]. Comparatively, we incorporate the *Frobenius* norm as a smoother version of sparsity, which can further preserve the intrinsic relationships of patches. Another benefit from $\|\cdot\|_F$ regularization is that it can ensure an analytical solution, therefore improving numerical stability and computational scalability. Overall, the affinity constraint and $\|\cdot\|_F$ regularization aim to consolidate similarity relationships among keypoints in a mutual reinforced way. Finally, SDFR could generate feature codes by considering both the discriminative capability and reconstruction accuracy, which is suitable to explore distinctive and diverse keypoints.

3.2. Optimization

We propose an iterative optimization algorithm for solving SDFR. Although the problem (1) is not jointly convex over all variables $\{c_1, c_2, ..., c_N\}$, it is convex with respect to one of them with the others fixed. Therefore, the objective function can be decomposed into the following N sub-problems:

Without loss of generality, by assuming variables $\{c_j,\ j \neq i,\ j = 1, 2, ...N\}$ are fixed, the feature code c_i of the i-th patch can be updated. Since those terms that are independent of variable c_i can be removed from the objective, we can derive the derivative of Eq. (1) as follows:

$$(D^T D + \beta I)c_i + c_i\alpha\sum_{j\neq i}^{N} S_{ij} = D^T f_i + \alpha\sum_{j\neq i}^{N} c_j S_{ij} \quad (3)$$

therefore, the feature code of the i-th patch can be analytically updated as: $c_i = (D^T D + \beta I + \alpha K I)^{-1}(D^T f_i + \alpha M)$, where letting $K = \sum_{j\neq i}^{N} S_{ij}$, $M = \sum_{j\neq i}^{N} c_j S_{ij}$ for brevity.

In addition, the initialization procedure is performed by solving: $\min_{\hat{c_i}} \| f_i - D \cdot \hat{c_i} \|_F^2 + \lambda \| \hat{c_i} \|_F^2$, which has an explicit solution given by: $\hat{c_i} = (D^T D + \lambda I)^{-1} D^T f_i$, where $\hat{c_i}$ is the feature code of the i-th patch for initialization. Furthermore, due to the high computational cost of the affinity constraint of Eq. (1), SDFR searches for k-nearest neighbors in the visual space to speed up the optimization process.

3.3. Keypoint Measure and Refinement

Once the feature codes of all patches are obtained, we measure the response strengths of them by: $RS = \| c_i \|_1,\ i = 1, 2, ..., N$. Then all the patches are sorted in descending order by the RS measure. Since structural information could be reflected by the combined weights of dictionary atoms, we point out that the higher response for a patch, the more likely it is considered to be a keypoint. This formula provides a simple yet efficient manner of capturing keypoint candidates.

Considering the phenomenon that many feature points are gathered together in local regions, we generate self-centered quadratic bounding boxes with a side length of $2 \times n$ for each patch, then the non-maxima suppression (NMS) is utilized to refine these keypoint candidates.

4. EXPERIMENTS

4.1. Datasets and Setting

We conduct experiments to testify SDFR by comparing it with several state-of-the-art keypoint detectors on the Webcam [1] and EF [14] benchmarks. The images from the above datasets are highly challenging due to the drastic changes in illumination, viewpoint, and background clutter.

Regarding the implementation details of SDFR, we partition the input image into 8×8 patches, and the number of the dictionary atoms is set to 256. The trade-off parameters α, β and λ are respectively set as 0.1, 0.08, 0.08, which are chosen within the range $[10^{-2}, 10^{0}]$ according to the experimental evaluation. For k-nearest research, we adopt $k = 10$ in the proposed model. Moreover, to better localize keypoints when images have suffered from diverse transformations, we develop three variants of the basic SDFR detector as follows:
SDFR-M: A spatial pyramid (3-levels) is incorporated into SDFR, which constructs scales with a 1.2 factor apart.
SDFR-D: The SDFR detector with overlapped patch dividing strategy (2 pixels step for EF; accelerate the evaluation process for Webcam by relaxing to 4 pixels step).
SDFR-MD: Overlapped patches are fed into SDFR-M.

4.2. Experimental Results

We compare SDFR to several key handcrafted, state-of-the-art semi-handcrafted and learning based methods, including SIFT [7], SURF [8], KAZE [11], AKAZE [12], (HarLap, HesAff) [6], SFOP [10], SCK-C [4], (T-P24, T-CNN) [1], ConDet [15] and DTC [3]. Note that T-P24, T-CNN, ConDet and DTC fall into the category of CNNs based approaches, which are strong competitors on the benchmarks. To quantify the performance, two evaluation criteria: '*repeatability*' and '*matching score*' defined in [23, 24] are used to measure the precision of keypoint detection. We also fix 1000 keypoints per image following the common experimental setting [3, 4].

Table 1 reports the repeatability comparisons on the Webcam and EF datasets. It can be seen that SDFR clearly outperforms the best handcrafted and learning based detectors. In particular, it shows significant improvement over any of the compared methods by a large margin on Webcam. It is noteworthy that SDFR is performed in a completely unsupervised manner. This surprising result confirms the merits of SDFR from the previous section that (i) SDFR inherits the sparsity property from sparse coding for representing patches, thus enabling resistance against distractors, and (ii) the affinity constraint further consolidates the consistency of feature

codes for similar patches. The above two merits are well suited to where the images mainly undergo illumination changes (like Webcam), thereby easily localizing the repeatable keypoints. It is also interesting to notice that the overlapped version of SDFR hurts the repeatability. This is because there are no viewpoint and scale changes on Webcam, which likely stimulates the sensitivity of the detector to local distractors caused by overlapped patches.

Table 1. Repeatability (%) comparisons on Webcam and EF (best result bold, second best underlined).

Method	Webcam	EF	Method	Webcam	EF
SIFT [7]	29.5	20.8	AKAZE [12]	65.8	32.2
SURF [8]	46.0	39.7	SCK-C [4]	63.2	41.2
SFOP [10]	43.8	36.1	**SDFR**	**85.5**	44.1
HarLap [6]	48.2	35.7	**SDFR-M**	<u>84.0</u>	<u>49.6</u>
HesAff [6]	42.5	26.6	**SDFR-D**	80.6	47.9
KAZE [11]	55.7	30.1	**SDFR-MD**	80.9	**50.5**
Learning based					
T-CNN [1]	51.4	38.0	ConDet [15]	49.9	42.7
T-P24 [1]	61.7	45.4	DTC [3]	68.4	46.6

Furthermore, matching scores are reported in Table 2 to evaluate the performances in feature matching scenarios using detected keypoints. As we can see, SDFR achieves the best accuracies on the benchmark datasets. Here, we observe that the overlapped patches provide more accurate local spatial information for feature matching. Different from the repeatability degradation on Webcam, keypoints combined with feature descriptors may mitigate the sensitivity to trivial image distractors. Also, SDFR with the image pyramid facilitates the handling of EF since several sequences from this dataset have large changes in spatial resolutions. Moreover, we show the qualitative results in Figure 1, where SDFR is compared with several key handcrafted detectors like FAST [25], SFOP [10] and SIFT [7]. For fair comparisons, the number of keypoints per image is fixed to near 1000 and the SIFT descriptor is used to generate feature vectors for all the comparing detectors. The ratio and the number of correct matches for each image pair are shown at the bottom of the figures. From this comparison, the improved performance can be obtained using the extracted keypoints by the SDFR model.

Visual dictionaries have the capability to reconstruct images using a few atoms. Figure 2(a) shows some example atoms in the proposed method, where we can see that clear local structures are leveraged to explore potential keypoints. We further testify the sensitivity of SDFR with respect to the model parameters. Specifically, we jointly evaluate how parameters α, β affect the matching performance on the EF dataset. From Figure 2(b), it is easy to observe that the performance tends to be better with the value of α becoming larger,

Table 2. Matching Score (%) comparisons on the benchmark datasets (best result bold, second best underlined).

Method	Webcam	EF	Method	Webcam	EF
SIFT [7]	12.9	10.2	SCK-C [4]	24.7	8.8
HesAff [6]	13.8	5.4	**SDFR**	30.5	8.1
AKAZE [12]	28.6	6.5	**SDFR-M**	30.7	**10.9**
T-P24 [1]	13.4	5.2	**SDFR-D**	<u>32.2</u>	8.4
DTC [3]	19.4	6.2	**SDFR-MD**	**32.4**	<u>10.3</u>

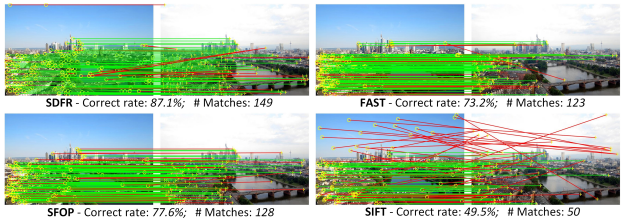

Fig. 1. Matching example of comparisons from Webcam, including SDFR, FAST [25], SFOP [10], and SIFT [7].

which could achieve its best result in the range of $[10^{-1}, 10^{1}]$. This verifies the effectiveness of the affinity constraint in the SDFR model. From the analysis on β, we can see that the performance increases slightly when enlarging β, but would degrade if β is set too large (≥ 100). A reasonable explanation is that imposing a large weight on the sparsity will make most entries of feature codes approach zero, thus leaving the discriminative power that is not fully explored.

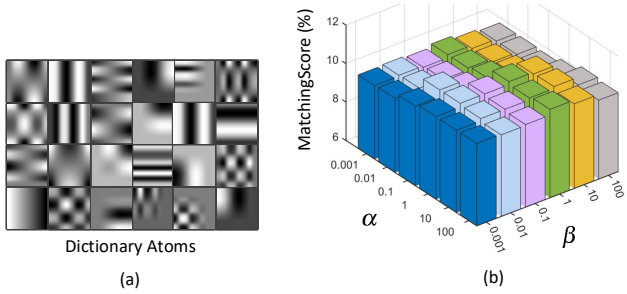

Fig. 2. (a) visualization of example atoms from the dictionary, and (b) parameter influence of α, β using the EF dataset.

5. CONCLUSION

In this work, we propose a keypoint detection method through optimizing the objective of a discriminative feature encoding. The main merit of the proposed model is that it is capable of characterizing the relationship of feature points for favorably discovering distinctive and diverse local structures implied in images. Experimental evaluations on benchmarks were testified to demonstrate its superior performance against baseline and state-of-the-art approaches.

6. REFERENCES

[1] Y. Verdie, Kwang Moo Yi, P. Fua, and V. Lepetit, "Tilde: A temporally invariant learned detector," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5279–5288.

[2] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.net: Keypoint detection by handcrafted and learned cnn filters," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 5835–5843.

[3] X. Zhang, F. X. Yu, S. Karaman, and S. Chang, "Learning discriminative and transformation covariant local feature detectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4923–4931.

[4] T. Hong-Phuoc and L. Guan, "A novel key-point detector based on sparse coding," *IEEE Transactions on Image Processing*, vol. 29, pp. 747–756, 2020.

[5] Christopher G. Harris and Mike Stephens, "A combined corner and edge detector.," in *Alvey Vision Conference*, 1988, pp. 1–6.

[6] Krystian Mikolajczyk and Cordelia Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, pp. 63–86, 2004.

[7] David Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[8] 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.

[9] Jiri Matas, Ondrej Chum, Martin Urban, and Tomas Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, pp. 761–767, 2004.

[10] W. Forstner, T. Dickscheid, and F. Schindler, "Detecting interpretable and accurate scale-invariant keypoints," in *IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2256–2263.

[11] Pablo Fernndez Alcantarill, Adrien Bartoli, and Andrew J. Davison, "Kaze features," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 214–227.

[12] P. F. Alcantarilla, J. Nuevo, and A. Bartoli, "Fast explicit diffusion for accelerated features in nonlinear scale spaces," in *British Machine Vision Conference (BMVC)*, 2013.

[13] Daniel Hauagge and Noah Snavely, "Image matching using local symmetry features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 206–213.

[14] C. L. Zitnick and K. Ramnath, "Edge foci interest points," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 359–366.

[15] Karel Lenc and Andrea Vedaldi, "Learning covariant feature detectors," in *ECCV Workshop on Geometry Meets Deep Learning*, 2016, pp. 100–117.

[16] T. Hong Phouc and L. Guan, "A scale and rotational invariant key-point detector based on sparse coding," *ACM Trans. on Intelligent Systems and Technology (accepted)*, 2021.

[17] Yurui Xie and Ling Guan, "Automatic sparsity-aware recognition for keypoint detection," in *IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 127–134.

[18] M. Aharon, M. Elad, and A Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[19] Georgi P. Tolstov and Richard A. Silverman, "Fourier series," in *Courier Corporation*, 1976.

[20] Karl Skretting and Kjersti Engan, "Learned dictionaries for sparse image representation: Properties and results," in *Wavelets and Sparsity XIV*, 2011, vol. 8138, pp. 404–417.

[21] S. Gao, I. W. Tsang, L. Chia, and P. Zhao, "Local features are not lonely - laplacian sparse coding for image classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3555–3561.

[22] K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1605–1612.

[23] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, pp. 43–72, 2005.

[24] K. Lenc, V. Gulshan, and A. Vedaldi, "Vlbenchmkars," http://www.vlfeat.org/benchmarks/, 2011.

[25] Edward Rosten and Tom Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision (ECCV)*, 2006, pp. 430–443.