Image Correspondences Matching Using Multiple Features Fusion

Song Wu, Michael S. Lew

The Leiden Institute of Advanced Computer Science,
Leiden University, Netherlands
{s.wu, m.s.lew}@liacs.leidenuniv.nl

Abstract. In this paper, we present a novel framework which significantly increases the accuracy of correspondences matching between two images under various image transformations. We first define a retina inspired patch-structure which mimics the human eye retina topology, and use the highly discriminative convolutional neural networks (CNNs) features to represent those patches. Then, we employ the conventional salient point methods to locate salient points, and finally, we fuse both the local descriptor of each salient point and the CNN feature from the local patch which the salient point belongs to. The evaluation results show the effectiveness of the proposed multiple features fusion (MFF) framework and that it improves the accuracy of leading approaches on two popular benchmark datasets.

Keywords: Salient points methods, convolutional neural networks, correspondences matching

1 Introduction

Vision applications such as 3D object reconstruction [1], image stitching [2] as well as object tracking in video sequences [3] [4] mainly rely on the correct correspondences matching across images.

The determination that whether a pair of salient points is correctly corresponding to each other is a quite challenging task. This is mainly due to the existing scale, rotation, and viewpoint transformations between the compared images. The past decades witnessed the effectiveness of salient point methods to this issue. The salient point methods firstly locate extrema (the candidate salient points) in the image scale space, and then generate a local descriptor to characterize each salient point. Finally, the nearest neighbor point obtained by the similarity measure is determined as the correspondence. A representative salient point method is SIFT [5], which detects the salient points in Difference-of-Gaussians scale space, and uses the orientation histogram of gradient to represent these obtained salient points. Most of other efforts (such as SURF [6], KAZE [7] were presented to improve the efficiency or accuracy of salient points localization. The SURF method makes use of a box-filter to approximate to the commonly used Laplace of Gaussian (LoG), and further employs the integral image to speed

Fig. 1. Two salient points share the same nearest neighbor point between the compared images. However, the green line is a false match, and the yellow line is defined as a correct match. This is because the similarity of yellow patch is better than green patch when compared to the blue patch.

up the box-filter based scale space construction. The recent KAZE employs a nonlinear scale space and combines with the Additive Operator Splitting (AOS) and special conductance diffusion to reduce noise. The nonlinear scale space could retain the object boundary structure and generate more accurate positions for salient points. Furthermore, the local binary representations were proposed with the advantages of fast computation and low memory requirements (BRIEF [8], ORB [9], BRISK [10], and FREAK [11]). The generation of local binary descriptors is mainly based on the pair-wise intensities comparison in a pre-defined structure. However, the local binary descriptors focus primarily on improving the speed and storage rather than the precision.

The goal of this paper is to improve the accuracy of correspondences matching from salient point methods. We propose a novel multiple feature fusion (MFF) framework in this paper and it shows the robustness to the challenging transformations (such as the rotation and perspective changes). Our framework is motivated by the theory of global precedence that humans perceive the global structure before the fine level local details. As illustrated in Fig. 1, the proposed framework combines the low-level local feature of salient point together with the high-level feature in its surrounding patch in the pre-defined global structure to establish the correct correspondences matching.

There are two important roles in the proposed framework: one is how to define the global structure in the image, and the other one is what kinds of features are appropriate to represent these patches in the pre-defined global structure. Specifically, we employ a retina inspired sampling pattern to construct a retina patch-structure in the image. The retina sampling pattern could effectively mimic the topology of the retina in human vision system. Moreover, inspired by the fact that the image representations built upon convolutional neural networks (CNNs) [12] have strong discrimination, we choose to describe these patches via high-level CNN features. The performance evaluation on two popular benchmark datasets demonstrated that the proposed MFF framework could significantly increase the

accuracy, stability, and reliability of correspondences matching under various image transformations, especially for the rotation and perspective changes.

The rest of the paper is organized as follows: Section 2 gives a brief review of related works. The construction of the proposed MFF framework is presented in Section 3. In Section 4, we describe the datasets and evaluation criterion in the experiment. The performance results of the MFF are shown in Section 5, and conclusions are given in Section 6.

2 RELATED WORK

Because of the high performance of deep convolutional neural networks in various computer vision applications, the CNNs based image correspondences matching is receiving increasing attention. Fischer et al. [13] extracted salient regions in an image via MSER detector. The extracted regions were normalized to a fixed resolution and then passed through a pre-trained convolutional neural network, and the output of the last layer in the CNN is used to represent the patch. Long et al. [14] and Tulsiani el al. [15] proposed to predict the salient points based on the convnet features from the output of CNN architecture. The recent methods mainly focus on the supervised learning schemes. Zagoruyko et al. [16] and Han et al. [17] used a Siamese network architecture which minimizes a pair-wise similarity loss of annotated pairs of raw image patches to jointly learn the features of local patches as well as the similarity metric for these local patches. The triplet network [18] employs the triplet ranking loss which can preserve the relative similarity relations of learn features to represent local patches. The framework introduced in this paper fuses the low-level local feature from each salient point and the high-level CNN feature from the patch it belongs to in order to achieve accurate correspondences matching.

3 MULTIPLE FEATURE FUSION FRAMEWORK

3.1 Retina Sampling Pattern Review

The retina sampling pattern has been widely used in various computer vision applications [11] [19], and those approaches made good use of the topology of human retina inspired by neuro-biology research. The topology of human retina reveals that the spatial distribution density of cone cells in the human retina decreases exponentially with the distance metric from the center of retina. As the illustration of the cones density in Fig. 2 (a), our approach employed the similar retina topology to define the patches structure in the image. As shown in Fig. 2 (b), different size of blocks are placed at the image domain with high sampling density in the center area. The advantages of the proposed retina patch-structure are as follows: small numbers of patches (43 patches) cover almost all image domain which offers a good trade-off between accuracy and efficiency towards to the CNN features extraction; the size of the block is calculated respected to the log-polar and high density patches in the center image domain

such that more details could be captured in the center area. Additionally, the overlapping between two patches in the retina pattern structure aims to increase the matching performance.

(a) (b)

Fig. 2. (a) The illustration of the density distribution of cones in the human retina. (b) The retina patch-structure in the image domain.

3.2 High Level Feature from CNNs

The motivation to utilize the output from a pre-trained CNN to represent the retina patch-structure stems from several properties of CNN features: First, the discrimination power of a CNN feature is significantly high and it outperforms those manually designed features in various computer vision applications by a large margin; Second, the CNN features are transferrable: some projects [20] [21] have demonstrated that the pre-trained networks still work well when they are applied to other vision tasks different from the datasets they were trained on.

The performance evaluation is based on a popular network architecture p-resented in Krizhevsky et al. [12] (AlexNet), which was trained on 1.2 million images from the ILSVRC2012 for classification (note that other high performance networks such as VGGNet, GoogleNet can also be used in the proposed framework.). The AlexNet network architecture consists of five stacked convolutional layers followed by normalization layers and pooling layers as well as two fully connected layers and a softmax classifier on top. Each fully connected layer contains 4096 neurons, and we use the CAFFE implementation [22] to extract the activations from the last two fully connected layers to represent each patch (referred to as fc6, fc7 and with the dimensional of 4096, respectively).

3.3 Multiple Features Fusion Framework

Towards to the local features of salient points and CNN features from the patches in the retina sampling pattern, we propose a novel feature fusion framework. For a specific salient point $P(x, y)$ in image I, first, we calculate its local descriptor f,

which is invariant to scale, rotation, and noise (such as SIFT, SURF, etc.). Then we calculate the distance between salient point position and the center of each retina patch to determine which patch the salient point belongs to. Finally each salient point is assigned a feature set: $F_{P(x,y)} = \{f, fc6_i, fc7_i\}$, where $i \in N$, which means $P(x, y)$ belongs to the ith patch in the retina patch-structure and N is the total amount of retina patches.

As the large variation in the value distribution from the directly obtained CNN features, the normalization operation is necessary. Inspired by the normalization of rootSIFT [23] which is more distinctive than SIFT, we apply the same normalization to the original CNN features, which exerts the feature vectors $L1$ normalization and then square root.

We define the similarity measure $S(P, P')$ to determine if two salient points $P(x, y)$, and $P'(x', y')$ is a correspondence as following:

$$S(P, P') = \exp(s(f, f')) \times (s(fc6_i, fc6'_j) + s(fc7_i, fc7'_j)) \tag{1}$$

where the $s(.)$ denotes the Euclidean metric, and we use an exponential function in order to emphasize the distance of two salient local descriptors.

Moreover, taking into the consideration that the existing overlaps in the proposed retina patch-structure, we use multiple assignment (MA) strategy to each salient point, which means that each salient point will be assigned K CNN features from its K nearest patches centers, and the similarity measure $S(P, P')$ is then updated as:

$$S(P, P') = \exp(s(f, f')) \times \sum_{i,j=1}^{K} (s(fc6_i, fc6'_j) + s(fc7_i, fc7'_j)) \tag{2}$$

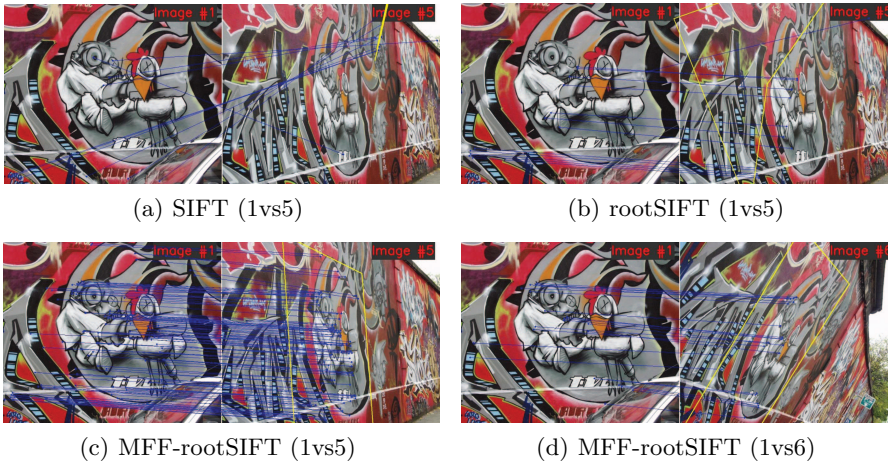
The performance of correspondences matching in Fig. 3 demonstrated the strength of our MFF in the cases of challenging perspective transformations in comparison to the popular SIFT, and rootSIFT.

4 EXPERIMENT SETUP

In this section, we conduct experiments to show the effectiveness of the proposed MFF framework. The accuracy of correspondences matching is evaluated on the MFF-rootSIFT and MFF-SURF, which applied our novel framework and compared with the leading popular approaches: SIFT, SURF, rootSIFT. The experimental environment for the evaluation is: Intel quad Core i7 Processor (2.6GHz), 12GB of RAM, and NVIDIA GTX970 with 4GRAM. The parameters of each compared salient point methods were set to the defaults and our MFF implementation is available online at: http://press.liacs.nl/researchdownloads/.

4.1 Datasets

The evaluation of correspondences matching is performed on two benchmark datasets (Mikolajczyk and Schmid [24] and Fischer et al. [13]), which both



(a) SIFT (1vs5) (b) rootSIFT (1vs5)

(c) MFF-rootSIFT (1vs5) (d) MFF-rootSIFT (1vs6)

Fig. 3. Illustration of correspondence matching, the MFF is applied on rootSIFT and compared to SIFT, rootSIFT on challenge affine object detection (graffiti 1vs5, graffiti 1vs6 proposed by Mikolajczyk and Schmid [24]). Our framework exactly located the position of object after homography estimation by the RANSAC.

provided the ground-truth homography between the reference image and the transformed image. The first dataset contains eight groups, and each group consists of six image samples (total 48 images) with various transformations (rotation, viewpoint, scale, JPEG compression, illumination and image blur). Considering the small scale of the dataset offered by Mikolajczyk and Schmid [24], a large dataset provided by Fischer et al. [13] is also employed in the evaluation. It contains 16 groups and each group contains 26 images (total 416 images) which are generated synthetically by applying 6 types of transformations (zoom, blur, illumination, rotation, perspective and nonlinear).

4.2 Evaluation criterion

As MFF is a framework to increase the correspondences matching, we use the defined formula (2) to establish the correspondences. While for the compared salient point methods, KD-tree index is established and the Nearest Neighbor Distance Ratio ($NNDR$) is used as the matching strategy to find the similar descriptors. $NNDR$ defines that two points will be considered as a match if $\parallel D_A - D_B \parallel / \parallel D_A - D_C \parallel$, where D_B is the first and D_C is the second nearest neighbor to D_A. The $NNDR$ matching threshold is set to 0.8 in the experiment.

To further determine whether a match is correct or not, we enforce a one-to-one constraint so that a match is considered as a correct only if its matching point is geometrically the closest point within the defined pixel coordinate error. For two compared images I and I', let the set of all matches as:

$$M = \{p_i \leftrightarrow p'_j | m(p_i, p'_j)\} \tag{3}$$

where $m(p_i, p'_j)$ denotes the two matches satisfy the correspondence requirement. We need to note that different points in image I could be projected to the same point in image I (many-to-one matches), even though only one single best match is returned for each point in reference image, and then we refine them to one-to-one match by accepting only the p_i with the smallest distance measure.

$$M_{refine} = \{p_k \leftrightarrow p' \in M | k = \arg\min_i m(p_i, p')\} \tag{4}$$

and the final correct matches are evaluated by the ground-truth homography:

$$correct\_match = \{p_i \leftrightarrow p'_j | D(H(p_i), p'_j) < \varepsilon\} \tag{5}$$

where $D(H(p_i), p'_j)$ is the position error after the ground-truth homography H projection for the point in image I, and in all cases, the ε is set to 3 pixels.

Following the common practice in evaluation protocols, we use the amount of correct matches as a criterion, which computes the total number of correct correspondence matches between two compared images.

5 Evaluation Results

In this section, we apply our MFF framework on the local features of SIFT, rootSIFT and SURF and present the detailed comparison performance on two benchmark datasets.

Impact of multiple assignment size: We first analyse the impact of the size of MA. Table 1 shows that the increasing size of MA marginally improves the performance of matching accuracy on both datasets. As a large value of MA size accordingly introduces noise, the value of MA size is set to 2 in the experiment.

MA size	Accuracy on dataset [24]		Accuracy on dataset [13]	
	MFF-SIFT	MFF-SURF	MFF-SIFT	MFF-SURF
1	575	381	2181	889
2	589	410	2344	990
3	588	408	2354	990

Table 1. The average number of correct matches under different MA size settings.

Evaluation results: We first evaluate the performance of each method on the dataset proposed by Mikolajczyk and Schmid [24]. The number of correct matches and the results under perspective, scale and rotation changes are shown in Fig. 4, and they clearly illustrates the effectiveness of our proposed MFF framework. Note that the MFF-rootSIFT obtained the highest number of correct matches in all cases, and MFF-SURF also obtained better performance than original SURF method.

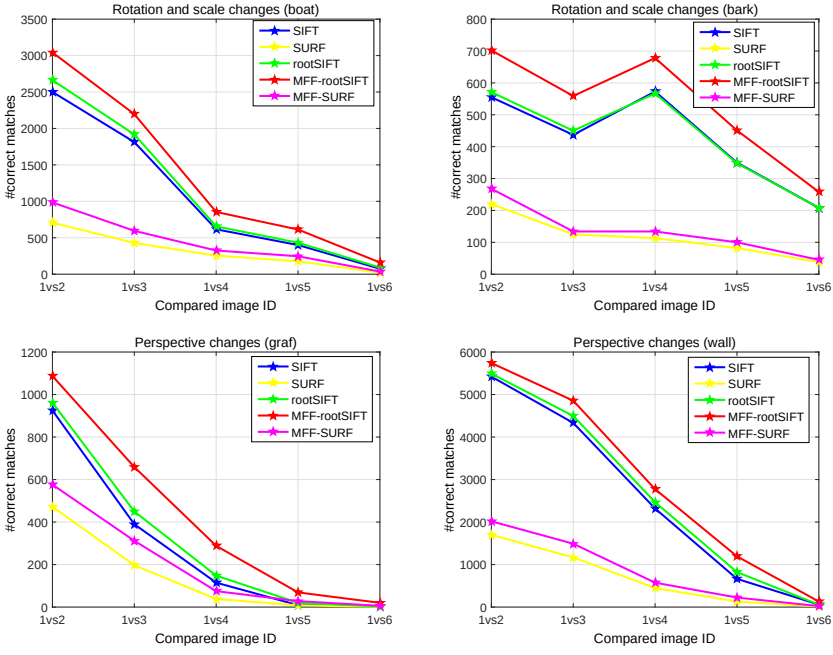


Fig. 4. Evaluation results on the viewpoint, rotation and scale changes based on the dataset provided by Mikolajczyk and Schmid [24].

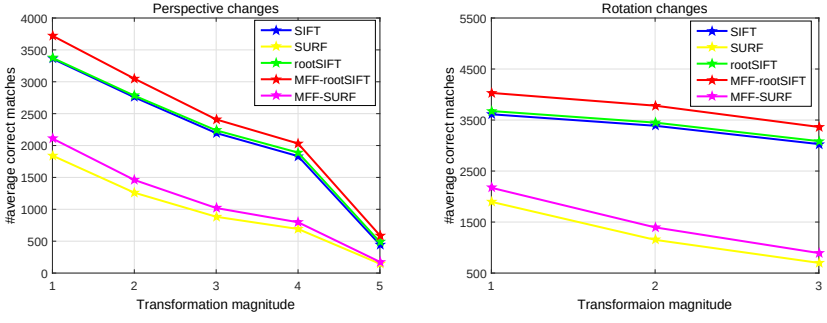


Fig. 5. Evaluation results on the viewpoint and rotation transformation based on the dataset of Fischer et al. [13].

We then evaluate all these approaches on a large scale dataset designed by Fischer et al. [13]. We use the average score of correct matches to measure the performance, and the evaluation results under two challenging transformations of viewpoint and rotation are shown in Fig. 5. It can be observed that similar tendency are demonstrated compared to the results illustrated in Fig. 4, and this further demonstrates that the MFF can significantly increase the matching

accuracy under various transformations. The evaluation results in Fig. 4 and 5 both show that the proposed MFF framework is effective and can significantly improve the accuracy of correspondences matching when combined with the traditional salient point methods.

6 Conclusions

This paper propose a novel MFF framework. It firstly computes a retina inspired patch-structure and locates the salient points in an image. Then the MFF fuses the local descriptor of each salient point and the CNN feature extracted from the patch around the salient point. The experimental results demonstrate the effectiveness of the proposed framework and it yields higher accuracy in correspondences matching under the viewpoint, scale and rotation changes.

Acknowledgments. We are grateful to the support of NVIDIA for this work.

References

1. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. International Journal of Computer Vision **66**(3) (2006) 231–259
2. Yang, X., Wang, M.: Seamless image stitching method based on asift. Comput Eng **39**(2) (2013) 241–244
3. Yang, H., Shao, L., Zheng, F., Wang, L., Song, Z.: Recent advances and trends in visual tracking: A review. Neurocomputing **74**(18) (2011) 3823–3831
4. Gauglitz, S., Höllerer, T., Turk, M.: Evaluation of interest point detectors and feature descriptors for visual tracking. International journal of computer vision **94**(3) (2011) 335–360
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2) (2004) 91–110
6. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3) (2008) 346–359
7. Alcantarilla, P.F., Bartoli, A., Davison, A.J.: Kaze features. In: Computer Vision–ECCV 2012. Springer (2012) 214–227
8. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: Brief: Binary robust independent elementary features. Computer Vision–ECCV 2010 (2010) 778–792
9. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: an efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 2564–2571
10. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 2548–2555
11. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, Ieee (2012) 510–517

12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
13. Fischer, P., Dosovitskiy, A., Brox, T.: Descriptor matching with convolutional neural networks: a comparison to sift. arXiv preprint arXiv:1405.5769 (2014)
14. Long, J.L., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: Advances in Neural Information Processing Systems. (2014) 1601–1609
15. Tulsiani, S., Malik, J.: Viewpoints and keypoints. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2015) 1510–1519
16. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4353–4361
17. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying feature and metric learning for patch-based matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3279–3286
18. Kumar, B., Carneiro, G., Reid, I.: Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. arXiv preprint arXiv:1512.09272 (2015)
19. Wu, S., Lew, M.S.: Riff: Retina-inspired invariant fast feature descriptor. In: Proceedings of the ACM International Conference on Multimedia, ACM (2014) 1129–1132
20. Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2014) 806–813
21. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. (2014) 3320–3328
22. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, ACM (2014) 675–678
23. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2911–2918
24. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(10) (2005) 1615–1630