Image Registration Based on Patch Matching Using a Novel Convolutional Descriptor

Wang Xie, Hongxia Gao(✉), and Zhanhong Chen

School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, Guangdong, People's Republic of China
hxgao@scut.edu.cn

Abstract. In this paper we introduce a novel feature descriptor based on deep learning that trains a model to match the patches of images on scenes captured under different viewpoints and lighting conditions. The patch matching of images capturing the same scene in varied circumstances and diverse manners is challenging. Our approach is influenced by recent success of CNNs in classification tasks. We develop a model which maps the raw image patch to a low dimensional feature vector. As our experiments show, the proposed approach is much better than state-of-the-art descriptors and can be considered as a direct replacement of SURF. The results confirm that these techniques further improve the performance of the proposed descriptor. Then we propose an improved Random Sample Consensus algorithm for removing false matching points. Finally, we show that our neural network based image descriptor for image patch matching outperforms state-of-the-art methods on a number of benchmark datasets and can be used for image registration with high quality.

Keywords: Feature descriptor · Deep learning · Patch matching

1 Introduction

Finding correspondences between image patches is one of the most widely studied issues in computer vision. Many of the most widely used approaches, like SIFT [1] or SURF [2] descriptors which have made a critical and wide impact in various computer vision tasks, are based on hand-crafted features and have limited ability to deal with negative factors such as noise which makes a search of similar patches more difficult. Recently, a variety of methods based on supervised machine learning have been successfully applied for learning patch descriptors which are always low dimensional feature vectors [3–5]. These methods are significantly superior to the hand-crafted approaches and promote our research in learned feature descriptors.

The discussion about comparison between learned feature descriptors and traditional handcrafted feature descriptors never stops. The deep feature has achieved the superior performance for many classification tasks, even fine-grained object recognition. While the performance improvements with CNN based descriptors come at the cost of extensive training time. Another issue in the area of matching patches is the limited benchmark data. The handcrafted local feature has been a subject of study in computer vision for almost twenty years, the recent progress in deep neural network

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 285–296, 2018.
https://doi.org/10.1007/978-3-030-03335-4_25

has led to a particular interest-learnable local feature descriptor. Specially, the features from the trained model of a convolutional network on ImageNet [12] can improve over SIFT in [9]. [10, 11] train a siamese deep network with hinge loss which have created great improvements in image patch matching (Fig. 1).

Fig. 1. We propose a new method for jointly learning key-point detection and patch-based representations in depth images towards the key-point matching objective.

The strategies of our novel feature descriptor learning are as follows: Our descriptor include feature point detector, orientation estimation and descriptor three parts, During the training phase, we use the image patches centroids and orientations of the key-points used by the Structure-from-Motion algorithm that we ran on images of a scene captured under distinct viewpoints and brightness to produce image patches. Siamese architecture is utilized to minimize a loss function with the similarity metric to be small for positive image patchpairs but large for negative image patchpairs. Then we conduct images registration with different viewpoints and illumination using our trained novel convolutional descriptor. We measure the key-point similarities by correlation of descriptors and we perform the final transformation by a new variant of Random-Sample-Consensus (RANSAC). As our experiments show, this new approach produces accurate registration results on images with different viewpoints and illumination settings.

In this paper we propose a descriptor based on CNN whose convolutional filters are learned to robustly detect feature points in spite of lighting and viewpoint changes. More over, we also use deep learning-based approach to predict stable orientations. Lastly, the model extract features directly from raw image patches with CNNs trained on large volumes of data. Those improve the performance of traditional hand-crafted method and has reduced matching error and increased registration accuracy.

The rest of the paper is organized as follows. In Sect. 2, we present related work focusing on patch matching problem and image registration. Section 3 describes the proposed method. In Sect. 4, we discuss implementation details and our experimental results, respectively. We provide conclusions in Sect. 5.

2 Related Work

Image registration via patch matching always revolves about matching the selecting feature descriptor and removing mismatched points via a Random-Sample Consensus algorithm to calculate the transform model. In this section, we will therefore discuss these two elements separately.

2.1 Feature Descriptors

Feature descriptors which are robust to transformations such as viewpoint or illumination changes have been widely applied for finding similar and dissimilar image patches in computer vision tasks. The feature descriptors are carefully designed from general measurement methods such as moment invariants, histograms of gradients in the past few years. SIFT [1] is computed from local histograms of gradient orientations and is distinguishable. However, the matching procedure is time-consuming owing to that the dimension of feature vector is high. Therefore, SURF [2] uses a low-dimensional vector representations to speed up the computation.

Nowadays, the trend has alternated from manually-designed methods to learned descriptors. Specially, end-to-end learning of patch descriptors using CNN has been developed in several works [9–11] and are far well compared to the state-of-the-art descriptors. It was demonstrated in [9] that the features from the trained model of a convolutional network on ImageNet [12] can improve over SIFT. Additionally, training a siamese deep network with hinge loss in [10, 11] based on positive and negative patch pairs, create vital improvements in matching achievement.

2.2 Image Registration

Image registration is useful in studying computer vision tasks such as getting the ultimate information from a combination of a great deal of divergent origins catching the same information in diverse circumstances and various manners and there are a great number of related literatures. Image registration methods [13, 14] perform an important part in scores of applications like image fusion. Early methods solve registration based on the gradients of the image such as [15]. Developed methods are using key-points [16, 17] and invariant descriptors to capture the geometric alignment.

According to the style of image acquisition, the utilization of Image registration can be divided into the following categories.

Multi-view Analysis. Capture images of similar object or scenes from multiple viewpoints to gain a better representation of the scanned object or scene. Examples include mosaicing of images and shape recovery from the stereo.

Multi-temporal Analysis. Images of the same scene are captured at different times usually under various conditions to notice alternations in the spectacle which emerge between the consecutive images acquisitions. Examples include motion tracking.

Multi-modal Analysis. Acquiring the images of the same spectacle via different sensors to merge the information obtained from a variety of sources to gain the minutiae of the spectacle.

An Image Registration task includes key-point detection, patch matching, conversion model assessment, image transformation determined.

3 Method

In this section, we first develop the complete feature descriptor. Then, So as to get the global transformation between the feature points, we introduce an approach which is an iterative RANSAC method to remove error matching from the same information in varied circumstances or diverse viewpoints after matching feature points.

3.1 Our Network Architecture

We select Faster R-CNN [8] with shared weights as the foundation for our network architecture due to that it is trained for the work of target detection and can offer us block representations and a trainable methods for choosing those patches. Then, image patches are linked to our ORI-EST network to predict stable orientations. After the image blocks has been rotated, image patches of both branches are connected to a fully connected layer to extract the feature vectors (Fig. 2).

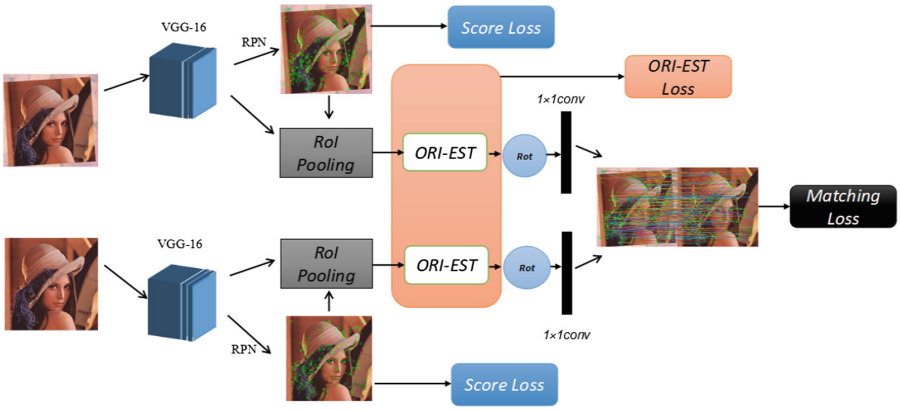

Fig. 2. Overview of our siamese architecture. Each branch uses VGG-16 as the base representation network. Features from conv5_3 are fed into both the Region Proposal network (RPN) and the region of interest (RoI) pooling layer, while their RoIs are fed to the RoI pooling layer, ORI-EST network and a fully connected layer to extract the feature vectors.

3.2 Descriptor

The descriptor can be formalized simply as

$$d = h_\rho(\boldsymbol{p}_\theta), \tag{1}$$

where $h(.)$ denotes the descriptor convolutional neural network, ρ its parameters, and $\boldsymbol{p}_\theta$ is the rotated patch from the Orientation Estimator. During the training phase, we use the image patches centroids and orientations of the key-points used by the Structure-from-Motion (SfM) algorithm to produce image patches $\boldsymbol{p}_\theta$.

To optimize the proposed network, we have to use a loss function which is able to discriminate positive and negative image patch pairs. More specifically, we train the weights of the network by using a loss function which prompts similar examples to be close, and dissimilar pairs to have Euclidean distance larger or equal to a margin m from each other.

$$L_{MatchLoss}(P_1, P_2, l) = \frac{1}{2N_{pos}} \sum_{i=1}^{N} lD^2 + \frac{1}{2N_{neg}} \sum_{i=1}^{N} (1-l)\{\max(0, m-D)\}^2, \tag{2}$$

where N_{pos} is the number of positive and negative pairs are represented by N_{neg} ($N = N_{pos} + N_{neg}$), l is a binary label is a positive ($l = 1$) or negative ($l = 0$) for choosing whether the input pair consisting of patch P_1 and P_2,, $m > 0$ is the margin for negative pairs and $D = \|h(P_1) - h(P_2)\|$ is the Euclidean Distance between feature vectors $h(P_1)$ and $h(P_2)$ of input images P_1 and P_2.

3.3 Orientation Estimation

SIFT determines the main orientation based on the histograms of gradient direction. SURF uses Haar-wavelet responses of sample points to extract the dominant orientation in the neighborhood of feature points.

We give a new orientation estimation approach for image patches. First, we introduce our convolutional neural networks then show details of our model. Given a patch $\boldsymbol{p}$ from the region computed by the detector, the Orientation Estimator estimates an orientation

$$\theta = f_w(\boldsymbol{p}), \tag{3}$$

where f denotes the Orientation Estimator CNN, and w its parameters.

We minimize a loss function $\sum_i L_i$ over the parameters w of a CNN, with

$$L_{ORI-ESTLoss}(\mathbf{p_i}) = \left\| h_\rho\left(\boldsymbol{p_i^1}, f_w\left(\boldsymbol{p_i^1}\right)\right) - h_\rho\left(\boldsymbol{p_i^2}, f_w\left(\boldsymbol{p_i^2}\right)\right)\right\|_2^2, \tag{4}$$

where the pairs $\boldsymbol{p_i} = \left\{\boldsymbol{p_i^1}, \boldsymbol{p_i^2}\right\}$ are pairs of image patches from the training dataset, $f_w\left(\boldsymbol{p_i^*}\right)$ means the orientation computed for image patch $\boldsymbol{p_i^*}$ using a CNN with parameters w, and $h\left(\boldsymbol{p_i^*}, \boldsymbol{\theta_i^*}\right)$ is the descriptor for patch $\boldsymbol{p_i^*}$ and orientation $\boldsymbol{\theta_i^*}$.

3.4 Feature Point Detectors

Each Faster R-CNN branch has a novel score loss for training the key-point detection stage, which is an uncomplicated but valid mean to recognize possibly stable key-points in training images. The score loss fine-tunes the parameters of the Region Proposal Network (RPN) of the Faster R-CNN [8] to obtain high-scoring proposals in regions of the image maps. We then use them to generate a score map whose values are local maxima at these positions. The region **S** proposed by the detector for patch **P** is computed as:

$$\boldsymbol{S} = g_{\mu}(\boldsymbol{p}), \tag{5}$$

where $g_{\mu}(\boldsymbol{p})$ denotes the detector itself with parameters μ

$$L_s(s,l) = \frac{1}{1+N_{pos}} - \frac{\gamma \sum_{i=1}^{N} l_{i\,log\,S_i}}{1+N_{pos}}, \tag{6}$$

where l_i is the label for the i^{th} key-point from image I whose value depends whether the key-point belongs to a positive or negative pair, S is the score of the key-point and γ is a regularization parameter.

$$L_{ScoreLoss}(\mathbf{p_i}) = \left\| h_{\rho}\left(\boldsymbol{p_i^1}, f_w\left(g_{\mu}\left(\boldsymbol{p_i^1}\right)\right)\right) - h_{\rho}\left(\boldsymbol{p_i^2}, f_w\left(g_{\mu}\left(\boldsymbol{p_i^2}\right)\right)\right) \right\|_2^2 + \lambda L_S(s,l), \tag{7}$$

λ is a regularization parameter.

3.5 Image Registration

Image registration is the procedure of aligning two or more images of the same scene which are captured from various sensors at different times or at multiple view-points. Image registration is significant in getting a better map of any alteration of a scene or object over a long time.

It is unavailable to use the group of all matches M to compute the final global transformation T between the images I_0 and I_1 in that a majority of matches in M are outliers. Therefore, it is necessary to apply RANSAC [18] for rejecting outliers before compute the transformation T. Moreover, In order to improve the accuracy of the transformation, we form the transformation T by our iterative RANSAC outliers rejection approach (Fig. 3).

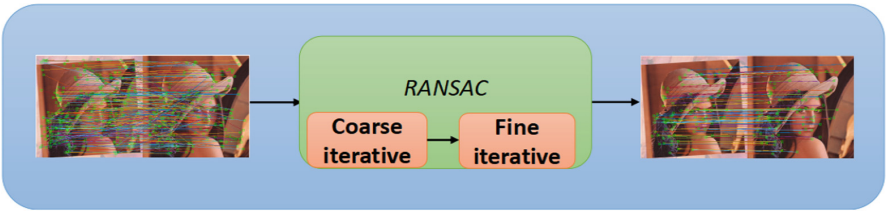

Fig. 3. Overview of RANSAC process, we propose a new RANSAC method for removing error key-point matching which is consisted of coarse and fine iterative.

The methods of iterative RANSAC are consisted of coarse iteration and fine iteration. The coarse iteration use RANSAC in a conventional way. We get a group of matches M_c by computing for each key-point $p \in I_0$ its best match $q^* \in I_1$. Obviously, this group includes inlier and outlier matches. The RANSAC outliers rejection approach is as follows, we sample subgroups of matches $m_1, ..., m_l, ... \in M_c$ and compute via least square the transformation T_l that most adapts these matches to each subgroup m_l. Therefore if our transformation T is characterized by n parameters, then we have $|m_l| = \lceil \frac{n}{2} \rceil$ since each match induces two linear constraints.

Ultimately, we choose T^* derived from the best group of matches m^* as the best transformation which has the greatest agreement in other matches. The number of other matches is formalized as $M_c - m^*$. A match agrees with a transformation if

$$\left\| T_{2\times 3} \begin{pmatrix} x_p \\ y_p \\ 1 \end{pmatrix} - \begin{pmatrix} x_{q^*} \\ y_{q^*} \\ 1 \end{pmatrix} \right\|_2 \leq RansacDistance, \tag{8}$$

the Ransac Distance in the first iteration is $r\, d_c$. T_c is expressed as the transformation that is found by RANSAC in the coarse iteration.

In the fine iteration we duplicate the same procedure as the coarse iteration, but use this initial guess T_c to limit the group of all matches in fine iteration M_f. More precisely, $p \in I_0$ can be matched to $q^* \in I_1$ only if their distance under T_c (like Eq. (8)), is less than MatchDistance. In fine iteration, MatchDistance $= md_f$ and RansacDistance $= rd_c$. We denote by T_f the transformation found by our fine iteration.

The parameters of the mapping function are computed with the established feature correspondence obtained from the previous step. Then, the mapping functions are applied for aligning the sensed image with the reference image.

4 Experimental Validation

In this section, we first present the datasets we used. We then present qualitative results, followed by a thorough quantitative comparison against a number of state-of-the-art baselines. The experiment was running on a machine with Ubuntu, Tensorflow, NVIDIA GeForce GTX 1080, Intel (R) Core (TM) i7-6700K CPU @ 4.00 GHz, 16 GB RAM. It took about one day to train our model. Our Input image size is about 2000 × 1000 and the runtime of testing process is about 12.5 s per image.

4.1 Dataset

We use the following two datasets to evaluate our method under illumination changes and multiple viewpoints, the *Webcam* dataset [6], which includes 710 images of 6 scenes with apparent illumination alternations but captured from the same viewpoint. The *Strecha* dataset [7], which involves 19 images of two scenes captured from manifest different viewpoints.

4.2 Qualitative Examples

We compare our method to the following combination of feature point detectors and descriptors, as reported by the authors of the corresponding papers: SIFT, SURF, ORB [19], PN-Net [20] with SIFT detector, and MatchNet [11] with SIFT detector.

A qualitative evaluation of the key-points shown in Fig. 4 reveals the tendency of the other methods to generate more key-points than ours. This demonstrates that our method is much less susceptible to the image noise, and validates our claim for learning the key-point generation process jointly with the representation.

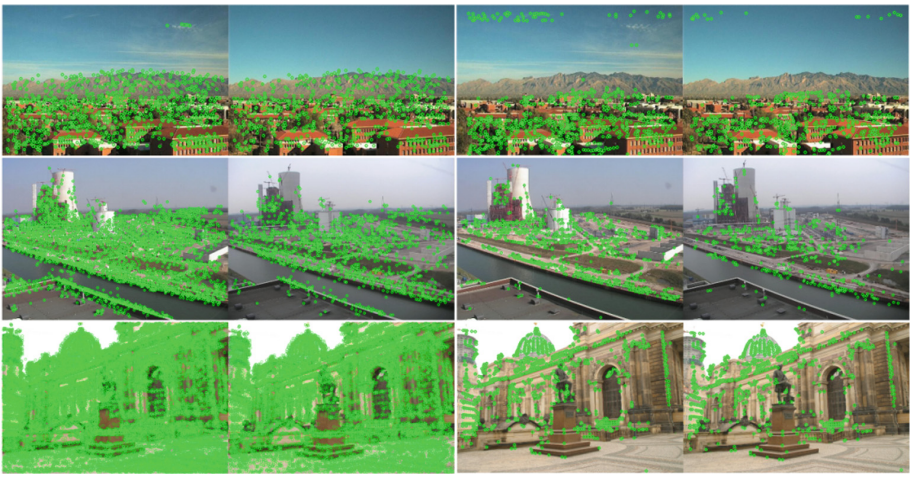

Fig. 4. Qualitative local feature matching examples of left: SURF and right: ours. Matches recovered by each method are shown in green color circles. SURF returns more key-points than ours. (Color figure online)

We compute the transformation T by RANSAC [18] rejection method. Figure 5 shows image key-points correct matching results, for both SURF and Ours. As expected, ours returns more correct correspondences.

These results show that our method outperforms traditional methods in matching correct key-points. Additionally, our method is much more reliable to the image under different conditions, and correct the mistakes of the original detectors.

4.3 Iterative RANSAC and Image Registration

The transformation T for every sample of matches from M is computed by least-squares. In order to ensure the accuracy of the transformation, we compute the transformation T by our iterative RANSAC outliers rejection method. Figure 6 shows image key-points correct matching results, for both RANSAC and our iterative RANSAC. As expected, ours returns more correct correspondences.

These results demonstrate that our method compares favorably with traditional RANSAC method in removing outliers.

Fig. 5. The figure shows the matching results after the traditional RANSAC. Feature matching examples of left: SURF and right: ours. Correct matches recovered by each method are shown in red color lines and the green color circles. Ours matches more key-points than SURF. (Color figure online)

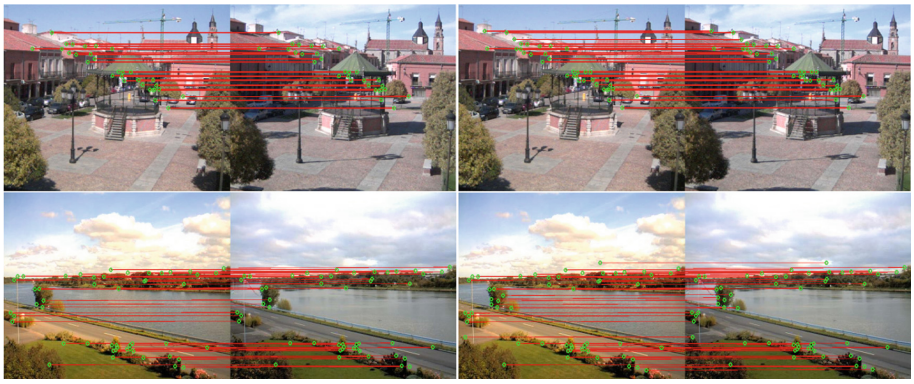

Fig. 6. The figure shows the matching results after the traditional RANSAC and our iterative RANSAC. Local feature matching examples of left: RANSAC and right: our iterative RANSAC. Matches recovered by each method are shown in red color lines and the descriptor support regions with green color circles. RANSAC matches less key-points than our iterative RANSAC matches. (Color figure online)

We use the *Webcam* dataset and the *Strecha* dataset to evaluate our method under illumination changes and multiple viewpoints. As our experiment show, most of the scenes are out door and with static objects but not include moving objects with a large obvious change in position. Our future work will focus on the registration for video frames under the scenes which are indoor and with some moving objects.

4.4 Quantitative Evaluation

In this section, we first present qualitative results, followed by a thorough quantitative comparison against a number of state-of-the-art feature descriptor baselines, which we consistently outperform. We then present our iterative RANSAC qualitative results, followed by traditional RANSAC (Fig. 7), (Tables 1 and 2).

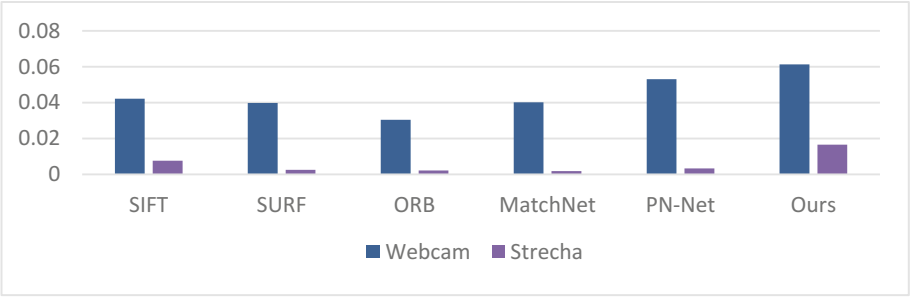

Fig. 7. Average matching score for all baselines.

Table 1. Average correct matching ratio for all baselines.

	SIFT	SURF	ORB	MatchNet	PN-Net	Ours
Webcam	.0422	.0398	.0304	.0402	.0531	.0613
Strecha	.0076	.0025	.0022	.0018	.0033	.0166

Table 2. Average correct matching ratio for different RANSAC.

	RANSAC	Our iterative RANSAC
Webcam	.0588	.0613
Strecha	.0157	.0166

5 Conclusions

We introduce a novel deep network architecture that combines the three components training a novel feature descriptor model to match the patches of images of a scene captured under different viewpoints and lighting conditions. The unified framework simultaneously learns a key-point detector, orientation estimator and view-invariant descriptor for key-point matching. Furthermore, we introduced a new score loss objective that maximizes the number of positive matches between images from two viewpoints. To remove false matching points, we propose an improved Random Sample Consensus algorithm.

Our experimental results demonstrate that our integrated method outperforms the state-of-the-art. A future performance improvement could be to study better structures of the orientation estimator network which could make the local feature descriptor even more robust to rotation transformations.

Acknowledgements. This work was supported by Natural Science Foundation of China under Grant 61603105, Fundamental Research Funds for the Central Universities under Grant 2015ZM128 and Science and Technology Program of Guangzhou, China under Grant (201707010054, 201704030072).

References

1. Lowe, D.: Distinctive image features from scale-invariant key-points. IJCV **60**(2), 91–110 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. CVIU **110** (3), 346–359 (2008)
3. Hua, G., Brown, M., Winder, S.: Discriminant learning of local image descriptors. IEEE Trans. Pattern Anal. Mach. Intell. (2010)
4. Trzcinski, T., Christoudias, C., Lepetit, V., Fua, P.: Learning image descriptors with the boosting-trick. In: NIPS, pp. 278–286 (2012)
5. Trzcinski, T., Christoudias, M., Fua, P., Lepetit, V.: Boosting binary key-point descriptors. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, Washington, DC, USA, pp. 2874–2881. IEEE Computer Society (2013)
6. Verdie, Y., Yi, K., Fua, P., Lepetit, V.: TILDE: a temporally invariant learned detector. In: CVPR (2015)
7. Strecha, C., Hansen, W., Van Gool, L., Fua, P., Thoennessen, U.: On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: CVPR (2008)
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, vol. 28, pp. 91–99 (2015)
9. Fischer, P., Dosovitskiy, A., Brox, T.: Descriptor matching with convolutional neural networks: a comparison to sift. Arxiv (2014)
10. Simo-Serra, E., Trulls, E., Ferraz, L., Kokkinos, I., Fua, P., Moreno-Noguer, F.: Discriminative learning of deep convolutional feature point descriptors. In: ICCV (2015)
11. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.: MatchNet: unifying feature and metric learning for patch-based matching. In: CVPR (2015)
12. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. IJCV 1–42 (2015)
13. Brown, L.: A survey of image registration techniques. ACM Comput. Surv. (CSUR) **24**(4), 325–376 (1992)
14. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)
15. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision (1981)
16. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, Manchester, UK, vol. 15 (1988). https://doi.org/10.5244/c.2.23
17. Lowe, D.: Distinctive image features from scale-invariant key-points. Int. J. Comput. Vis. **60**, 91–110 (2004)

18. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
19. Rublee, E., Rabaud, V., Konolidge, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: ICCV (2011)
20. Balntas, V., Johns, E., Tang, L., Mikolajczyk, K.: PN-Net: conjoined triple deep network for learning local image descriptors. arXiv Preprint (2016)