Wide Baseline Matching

Qianqian Wang
Cornell University
qw246@cornell.edu

Himank Yadav
Cornell University
hy539@cornell.edu

Wenqi Xian
Cornell University
wx97@cornell.edu

1. Problem statement

Finding correspondences between images is one of the most fundamental problems in computer vision, and a key component of the 3D reconstruction pipeline. While the problem of image correspondence has been studied extensively for decades, finding dense correspondences in a wide-baseline scenario still remains a challenging open problem. We aim to develop a deep architecture to solve the wide-baseline matching problem, i.e., the problem of establishing correspondences between a pair of images taken from very different viewpoints.

2. Related works

2.1. Traditional handcrafted methods

Considerable effort has been put into designing better features and descriptors to solve the wide-baseline matching problem. The Scale-invariant feature transform (SIFT) descriptor [9], which is invariant to rotation and scale, is one of the most widely used descriptors. The SIFT feature descriptor works well in certain cases for sparse correspondances but is not very helpful for dense wide-baseline matching. A variety of alternatives to the SIFT descriptor have been proposed, emphasizing speed (e.g, the Daisy descriptor [13]) or invariance to extreme transformations such as scale changes [7]. Though robust, the feature extraction, description as well as post-processing steps associated with the handcrafted descriptors require high-computational cost. Also, these handcrafted feature descriptors mainly leverage low-level image cues, lacking high-level semantic information.

2.2. Learning-based methods

Another line of work is using deep neural networks to establish correspondences between images. Some works try to estimate fundamental matrices and find inliers from an initial set of putative correspondences using neural networks [10][14]. Although these methods could achieve good results in the wide-baseline setting, they still rely on traditional feature extraction and matching pipelines, not in an end-to-end way. Recently, SuperPoint[4] uses fully convolutional networks to detect a richer set of interest points than the initial pre-adapted deep model and any other traditional corner detector. However, it still fails to extract enough keypoints in wide baseline scenarios. Another approach based on advances in deep neural networks to learn interest point detectors [5] shows that these learned detectors perform significantly better than hand-crafted features and offer an interesting insight that changing the receptive fields and striding effects beyond just using a good metric loss function has an impact on the quality of learned features. However, while this approach seems to do better than various previous metric learning techniques, the results still have some room for improvement while dealing with wide baseline scenarios.

3. Approach

We propose to use deep learning techniques for training visual descriptor, which learns to find dense correspondences for image matching. We start by constructing large amounts of wide-baseline correspondences from large-scale RGB-D dataset. We then train Siamese networks similar to Universal Correspondence Network [3], to directly learn the generic feature that preserves similarity for correspondences from pair of images. Specifically, we use correspondence contrastive loss and adopt hard-negative mining strategy to achieve better training efficiency. We find that combining the network with convolutional spatial transformer produces feature descriptors that are more robust to large viewpoint and scale changes. Experiments show that our network trained on wide-baseline correspondence identifies more correct keypoints than other pretrained image classification networks. Further experiments demonstrate that our learned descriptor can generalize well toward outdoor scenes, such as KITTI dataset [6].

3.1. Data Generation

We use Matterport3D dataset [2] which provides high-quality RGB-D images from a large variety of camera viewpoints in interior home environments. It enables us to construct large amount of wide-baseline correspondences required for training a strong model. As shown in Figure 1, training samples are selected from comprehensive cam-

era views and precisely computed based on camera poses, depth and surface normal annotations. We identified true correspondences if the estimated depth of the corresponding pixels falls within a radius of 10% of their true depth and if the inner product of their surface normals is larger than 0.3. We further removed correspondences that contain textureless regions using Harris Corner Detector.

Figure 1. Example training image pair. Red area highlights the matching correspondences.

3.2. Experiment Setup

We modify existing universal correspondence networks 2 for our task. Correspondence contrastive loss are used to supervise the end-to-end training process. Contrastive loss, as shown in the equation below, consists of two terms: the first minimizes the distance between positive pairs and the second pushes negative pairs in the embedding space to be away from each other by at least a margin of m. Correspondence contrastive loss is able to utilize more than 1k correspondences in one pair of training images, leading to faster training and convergence. Additionally, we use hard negative mining to mine negative pairs that violate the constrain the most and use those to expedite the training process. A negative pair is found by finding the nearest neighbor of extracted features from the first image that are far from the ground truth. We will discuss more implementation details and training process in the the following section.

$$L = \frac{1}{2N}\sum_{i}^{N}[s_i||F_{\mathcal{I}}(\mathbf{x}_i) - F_{\mathcal{I}'}(\mathbf{x}'_i)||^2 + (1-s_i)\max(0, m - ||F_{\mathcal{I}}(\mathbf{x}_i) - F_{\mathcal{I}'}(\mathbf{x}'_i)||)^2] \quad (1)$$

4. Experiments and Results

4.1. Implementation details

We adopt the Siamese network architecture [3] to learn the descriptors. For the feature representation extraction, We used the ImageNet pretrained GoogLeNet [12] from the bottom conv1 to the inception_4a layer, but stride 2 is used for the bottom 2 layers and 1 for the rest of the network. The ratio of positive and negative pairs for training is determined

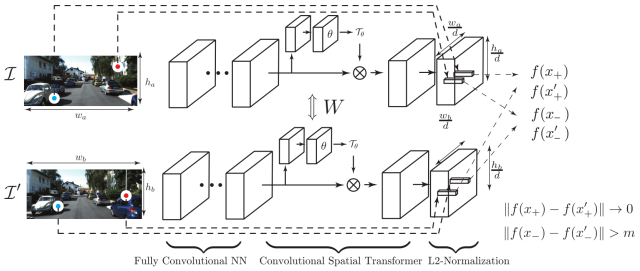

Figure 2. Overview of the universal correspondence network [3] architecture

in an adaptive manner, i.e., the more the hard negative examples found are, the larger the ratio of negative examples is. This strategy is found useful in helping the network escape from local minimum. However, due to the wide baseline nature of the dataset, at the very early stage the negative pairs will outnumber the positive pairs, thus dominating the training loss, which causes the network to easily converge to the local minimum by push all features away from each other. To address this problem, curriculum learning [1] is adopted, i.e., learning easier concepts first and gradually increase the difficulty. Here the difficulty is measured by the baseline. We train the dataset using image pairs with the baseline less than 3 meters and then gradually increase the baseline.

4.2. Evaluation Metrics

We use the percentage of correct keypoints (PCK) metric [11] to evaluate the matching performance. Specifically, for each feature in a query image, we find the nearest neighboring feature in the reference image as the predicted correspondence. If the predicted keypoint is closer than T pixels to the ground-truth keypoint, the correspondence is identified as correct. In contrast to many previous works, we apply no post-processing techniques, such as global optimization with an MRF. In this way, we are able to capture the performance of raw correspondences.

4.3. Experiments on spatial transformer

Convolutional neural networks can handle some degree of spatial transformations in images, i.e., moderate variances to scale and rotation. However, due to the nature of convolution, CNNs lack the ability to address large spatial transformations implicitly. Therefore, we adopt spatial transformer to explicitly enforce spatial information in the feature representation, which imitates the patch normalization in traditional handcrafted methods [9]. Unlike the global spatial transformer [8] which learns the affine transformation parameters for the whole image, we learn transformation parameters separately for each local region using the supervision of correspondence solely.

To justify the usefulness of spatial transformer, an ab-

lation study is done to see how the performance improves with the spatial transformer. We plot the PCK curves of the model trained with and without the spatial transformer in figure 3. It is shown that the spatial transformer significantly improves the PCK consistently across the whole range of threshold. Qualitative comparison is shown in figure 4, which can be seen that the correspondences are more accurate, though not precise, when utilizing the spatial transformer.

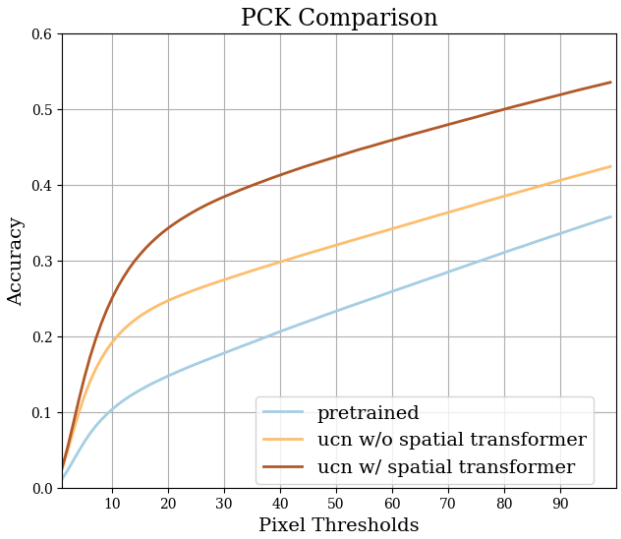

Figure 3. PCK comparison w.r.t spatial transformer

4.4. Properties of the learned descriptor

The ideal local image feature should achieve both distinctiveness and robustness. In order to match well, the feature descriptor should be invariant to the change of illumination, translation, scale and rotation, etc. Since the feature descriptors are constructed through convolution, the invariance to translation is automatically satisfied. In addition, the pooling layers in the fully convolutional neural network can implicitly improve the network's ability to handle minor changes of scale and rotation in images, whereas the spatial transformer layer can explicitly address larger spatial transformations in images.

We illustrate how the learned features can deal with the scale and viewpoint variance in Figure 6 and Figure 7, respectively. The model trained with spatial transformer performs better at locating keypoints, although it can be misled by repetitive patterns in some images.

Figure 4. qualitative comparison between correspondences generated with spatial transformer (top) and without spatial transformer (bottom). Colors indicate correspondences.

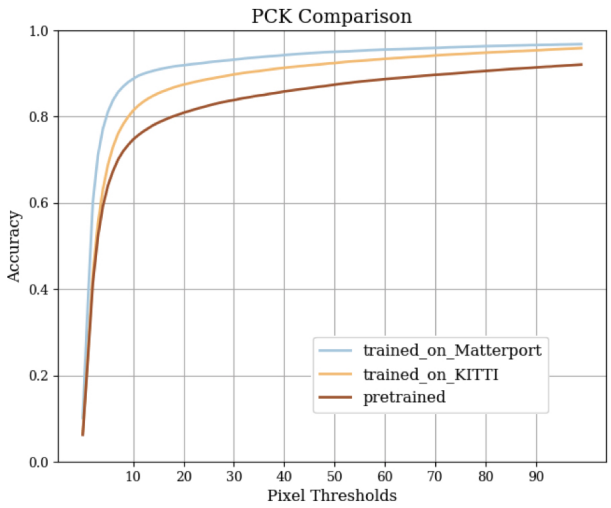

Figure 5. PCK evaluation on KITTI dataset

4.5. Generalization

We aim to learn generic descriptors which works well on all kinds of scenes in addition to indoor scenes. In other words, the descriptors learned on the Matterport3D dataset should be able to generalize well to arbitrary scenes. Although the network is fed with only indoor scene images, we find that it could generalize surprisingly well on outdoor scenes such as the KITTI [6] dataset. More specifically, the model trained without seeing any KITTI dataset even outperforms the one that trained with full supervision on the KITTI dataset. Figure 5 shows that the our trained model works better consistently through all pixel threshold. It's possible that the resistance to overfitting comes from the low level nature of the matching task, i.e., finding correspondences mainly relies on understanding of geometry which can be achieved without knowing extensive semantics in the image.

Figure 6. The scale invariance. The left, middle and right column are the nearest neighboring correspondences using feature representation from the GoogLeNet pretrained on ImageNet, the model learned without spatial transformer, and the model learned with spatial transformer respectively

Figure 7. The viewpoint invariance. The left, middle and right column are the nearest neighboring correspondences using feature representation from the GoogLeNet pretrained on ImageNet, the model learned without spatial transformer, and the model learned with spatial transformer respectively

References

[1] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 2

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[3] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. In *Advances in Neural Information Processing Systems*, pages 2414–2422, 2016. 1, 2

[4] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. *arXiv preprint arXiv:1712.07629*, 2017. 1

[5] M. E. Fathy, Q.-H. Tran, M. Z. Zia, P. Vernaza, and M. Chandraker. Hierarchical metric learning and matching for 2d and 3d geometric correspondences. *arXiv preprint arXiv:1803.07231*, 2018. 1

[6] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 3

[7] T. Hassner, V. Mayzels, and L. Zelnik-Manor. On sifts and their scales. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1522–1528. IEEE, 2012. 1

[8] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 2

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2

[10] R. Ranftl and V. Koltun. Deep fundamental matrix estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–299, 2018. 1

[11] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016. 2

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[13] 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*, 32(5):815–830, 2010. 1

[14] K. M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, number CONF, 2018. 1