Learning to Detect and Match Keypoints with Deep Architectures

Hani Altwaijry[1,2]
haltwaijry@cs.cornell.edu
Andreas Veit[1,2]
aveit@cs.cornell.edu
Serge Belongie[1,2]
sjb344@cornell.edu

[1] Cornell University
Ithaca, NY, USA

[2] Cornell Tech
New York, NY, USA

In computer vision, the extraction of effective features for the detection and description of important image regions is a key step for many applications. Traditionally, these features are extracted using hand engineered detectors and descriptors. Approaches adopting this paradigm are generally referred to as *keypoint-based* or *feature-based* approaches. Recently, the reintroduction of neural networks into many computer vision tasks broadly replaced hand-engineered feature-based approaches. Neural network based approaches generally learn the feature extraction as part an end-to-end pipeline. While these approaches have shown great success in tasks such as object detection and classification, other tasks such as structure-from-motion (SfM) still depend on purely engineered features, *e.g.* SIFT, to detect and describe keypoints.

In this paper, we propose a model that learns what constitutes a good keypoint, is capable of capturing keypoints at multiple scales and learns to decide whether two keypoints match. We achieve multiscale keypoint detection with a fully-convolutional network that recursively applies convolutions to regresses keypoint scores. With each successive convolution, the network evaluates image patches, i.e., keypoints, at a larger scale. By extracting the keypoint feature map after each convolution we obtain a feature map that resembles a keypoint scale-space. To learn descriptors for keypoint matching, we leverage a triplet network to learn an embedding where patches of matching keypoints are closer to each other than non-matching patches. Figure 1 provides an overview of our proposed model.

There is currently no large-scale dataset for learning both keypoint detectors and descriptors from image patches. Furthermore, finding training examples to train deep neural networks for this task poses a serious challenge, as collecting human annotated examples would be prohibitively expensive. Therefore, we create our own dataset by following a self-supervised approach. We utilize SfM to construct a large-scale model of 1.3 million 3D points, which are used to extract matching patches with varying photometric properties such as scale, illumination, perspective. Although those feature detections and matches were determined originally with engineered features, SfM factors in the underlying geometry. This allows to learn features that extend upon their engineered counterparts.

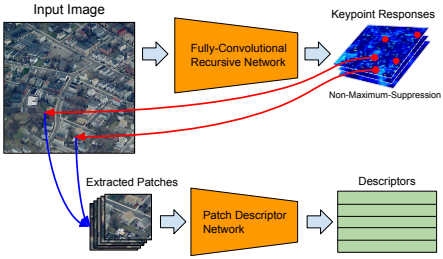

Figure 1: Proposed architecture for learning to detect and describe keypoints at multiple-scales.

We evaluate the proposed model both quantitatively and qualitatively and show its capablity of identifying multiscale keypoints as well as matching them. We show that the descriptors outperform previous approaches and demonstrate the transferability to unseen datasets with different statistics; Figure 2 shows an example.

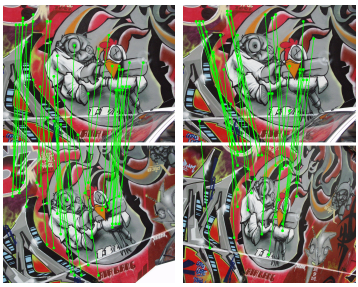

Figure 2: Qualitative evaluation on "Wall' image from the Oxford dataset.