Large scale visual search

Wu, S.

Chapter 6

Conclusions

6.1 Conclusions

In this thesis, we focus on large scale visual search. The topic of large scale visual search has seen a steady train of improvements in performance over the last decade. In this task, given a query image containing a specific object or scene, the goal is to return the images containing the same object or scene that may be captured from different viewpoints, under changed illumination and maybe occluded. The Bag-of-Words model was originally proposed for document retrieval. The introduction of salient point methods has made this model applicable to the image domain where it translates to the visual word model. General salient point methods involve a detector and a descriptor. The detector locates the salient regions in the image and the descriptor encodes discriminative information in the salient region into a local feature. Based on the salient point method, an image can be transformed into a collection of local feature vectors, which can be viewed as prototypes of words in text. The visual word model has been the state-of-the-art for many computer vision applications. It has greatly advanced the research of instance retrieval in the past ten years, and many improvements have been proposed.

One important aspect in the visual word model is the degree to which the salient point methods are invariant to image translation, scaling, and rotation, as well

as partially invariant to illumination changes, and robust to local geometric distortion. In Chapter 2, we presented a comparison of the existing salient point detectors and descriptors on diverse image distortions. These comparative experimental studies can benefit researchers in choosing an appropriate detector and descriptor for different computer vision applications. According to the evaluation results, we find that the FAST detector had the highest repeatability score compared to other detectors, moreover it had the lowest detection time-cost per point. Regarding the criterion of recall-precision, our experiments showed that the descriptors of SIFT, BRISK, and FREAK were the best performing affine invariant descriptors. Furthermore, evaluation of the time complexity showed that the binary descriptors are efficient with respect to feature description and matching.

Existing salient points methods tend to perform poorly to viewpoint changes. In Chapter 3, we presented the Retina-inspired Invariant Fast Feature, RIFF, which was designed for invariance to scale, rotation, and affine image deformations. The RIFF descriptor is based on pair-wise comparisons over a sampling pattern loosely based on the sampling pattern seen in the human retina and introduces a method for improving accuracy by maximizing the discriminatory power of the point set. The main contribution of the RIFF descriptor is in constructing the descriptor, where the discriminative power is optimized by ranking and deleting points with low distinctiveness. In our Bag-of-Words image retrieval tests on three well known datasets, RIFF outperformed the other feature descriptors with respect to invariance to scale, rotation, and affine transformations. Furthermore, we presented a performance evaluation of real valued and binary string salient point descriptors. The time complexity and space requirements showed that binary string descriptors are efficient in terms of feature extraction time and memory usage. With respect to the criterion of the mAP score, the image copy detection experiments showed some significant strength of binary string local descriptors: FREAK clearly outperformed SIFT on invariance to rotation, scale, and affine transformations; BRIEF had the best accuracy testing invariance to image blur and was among the best in robustness to cropping.

In recent years, the focus on image search has shifted from the visual word model to deep Convolutional Neural Networks (CNNs) features. The CNN is a hierarchical structure that has been shown to outperform hand-crafted features in a number of vision tasks, such as object detection, image segmentation, and classification. The power of CNNs mainly comes from the large number of parameters and the use of large scale datasets with rich annotations. Using the features extracted from CNN models, researchers have reported competitive performance compared to the classic visual word model. In Chapter 4, we proposed a novel image representation called deep binary codes which have important advantages over deep convolutional feature representations, as they can be calculated using a generic transferred model and therefore do not require additional training. The experimental results on well-known datasets as well as a large scale dataset show that deep binary codes are competitive to state-of-the-art approaches and can significantly reduce memory and computational costs for large scale image search. Moreover, in Chapter 5, we proposed to reuse the information in the previous layers in the network to recover the precision loss due to the pooling operation in the CNN. The presented Weighted Integration Architecture Network (WIAN) can enhance the power of the CNN model.

6.2 Future Work

In the future, we will try to improve our work in the following directions:

Convolutional neural networks based local descriptor generation: The generation of effective local image descriptors plays an important role in the applications of computer vision involving baseline stereo vision, structure from motion, visual words based image search, image classification and object detection, etc. The existing schemes of local descriptor generation can be categorized into hand-crafted or automatically learned schemes. Recent work focuses more on automatic learning of local descriptors. Learning based schemes usually optimize an objective function to generate robust and distinctive local descriptor. In particular, the most common objective functions are designed to minimize the

distance between the descriptors from the same 3D location (scale and location) or same class label extracted under varying imaging conditions and different viewpoints, and maximize the distance between patches from different 3D locations or different class labels. Concurrently, the automatically learning schemes of local descriptors based on deep convolutional neural networks have recently made dramatic progress. A Siamese network trained with a pair-wise loss ranking function and a triplet network trained with a triplet loss ranking function that also minimizes the distance (in the embedded space) between patches of the same labels and maximizes the distance between patches of different labels are used to automatically learn high performance local descriptors. However, all these methods suffer from huge training complexity, because they directly train CNNs using the pair-wise or triplet list, the length of which scales with the quadratic or cubic with the number of images in the training dataset. Therefore, it is important to further develop techniques to address huge training complexity while maintaining the robustness of the learned local descriptors. Another issue we need to address is the limitation of training data. The typical solution is to generate more training data from existing data using data augmentation schemes, such as scaling, rotating and cropping. Hence, it is important to further develop techniques for generating or collecting more comprehensive training data, which could make the networks learn better features that are robust to various changes, such as geometric transformations, and occlusion.

Convolutional neural networks based high level image representation: The outputs from the fully connected layer in the CNN are mostly used as image representation. However, the image representation from a fully connected layer suffers from the lack of description of local patterns, which is especially critical when occlusions or truncations exist in the images. With respect to the sensitivity to local stimulus, CNN features from the bottom or intermediate layers have shown promising performance. These discriminatively trained convolutional kernels respond to specific visual patterns that evolve from bottom to top layers. While capturing local activations, the intermediate features are less invariant to image translations. Compared to the pooling operation, which is usually utilized to map the intermediate features into global feature, one promising direction for

future research is to find more efficient ways to convert the intermediate features into low dimensional and high distinctiveness image representations, in order to avoid the information loss caused by pooling operations. Second, it is known that the top layers in CNNs are sensitive to semantics, while intermediate layers are specific to local patterns. For the image representation, we can obtain multiple layer features in the pre-trained CNN through one feed-forward step. It is not trivial to predict which layers are superior. Therefore, the fusion of the features from multiple layers is a good practice to further improve the accuracy of image search. Moreover, we can also fuse the features from different models to represent the image.

Convolutional neural networks based deep hash learning: In order to achieve efficient large scale image search, the high performance of the supervised deep hashing model appears to be promising. The first direction is to increase the ability to generalize by increasing the width or depth of the networks, for example, the width and depth of the CNN models in the literature [42, 51]. Larger networks could normally bring higher quality performance, but have the danger of over-fitting and require very large computational resources. A second direction is to define a good loss ranking function. As the commonly used pair-wise loss functions and triplet loss functions employ Euclidean distance to measure the similarity in the input space, we can replace the Euclidean distance with different similarity metrics for different input spaces. Moreover, we can also incorporate constraint information from the input space to the loss functions. A third direction towards more powerful models is to design more specific deep networks. Currently, almost all of the CNN-based schemes adopt a shared network for their predictions, which may not be distinctive enough. The study by Ouyang et al. [174] has verified that object-level annotation is superior to image-level annotation for object detection. This can be viewed as a kind of specific deep network that just focuses on the object region rather than the whole image. Another issue we need to note is that in some situations the amount of the annotated data is insufficient and it could result in over-fitting during the training of the CNN. Semi-supervised deep hashing makes use of the labeled data together with the

unlabeled data and may be able to overcome this limitation in the CNN training.