Large scale visual search

Wu, S.

English Summary

With the ever-growing amount of image data on the web, much attention has been devoted to large scale image search. It is one of the most challenging problems in computer vision for several reasons. First, it must address various appearance transformations such as changes in perspective, rotation and scale existing in the huge amount of image data. Second, it needs to minimize memory requirements and computational cost when generating image representations. Finally, it needs to construct an efficient index space and a suitable similarity measure to reduce the response time to the users. This thesis aims to provide robust image representations that are less sensitive to above mentioned appearance transformations and are suitable for large scale image retrieval.

Early approaches, the Bag-of-Words (BoW) model and its variants, have dominated the research on large scale image retrieval. The pipeline of BoW for image retrieval mainly consists of three steps: (i) salient point feature extraction; (ii) visual vocabulary generation; (iii) BoW based feature encoding. In each step, many efforts have been made to achieve state-of-the-art performance on large scale image search.

First, we investigated the strengths and weaknesses of the existing salient point detectors and descriptors on diverse image distortions. The comparative experimental studies we presented can support researchers in choosing an appropriate detector and descriptor to generate the BoW based image representation.

Compared to the real valued local descriptors, binary string local descriptors have the advantage of low memory requirements and efficient matching via Hamming distance. We further proposed to use the "K-majority" cluster method with ANN

search to generate a BoW image representation based on binary string local descriptors. The evaluation results showed that the binary string descriptor based BoW model has low memory requirements for vocabulary storage and competitive performance compared with real valued local descriptor based BoW image representation.

Since the existing salient point methods are sensitive to viewpoint or perspective changes, we further proposed a novel salient point descriptor named RIFF. RIFF is generated according to pair-wise intensity comparisons over a sampling pattern inspired by the human retina. The evaluation results showed that the RIFF based BoW image representations outperformed other feature descriptors with respect to invariance to scale, rotation, and viewpoint transformations.

More recently, image representations generated by the convolutional neural networks (CNNs) have demonstrated their high performance compared to the state-of-the-art for image retrieval. In this thesis, we explored both real valued and binary string image representations based on feature maps from the layers within CNNs. In addition, we presented a fusion scheme to further improve image search accuracy. Moreover, we designed a more powerful CNN architecture to improve the robustness of CNN models.

Finally, although this thesis makes a substantial number of contributions to large scale image retrieval, we also presented additional challenges and future research based on the contributions in this thesis.