A Deep Journey into Feature Extraction in Image Processing – A Survey

Mak Chun Chung, Desmond

The Hong Kong Polytechnic University

Abstract

Feature extraction is one of the important roles in the computer vision area. For instance, image classification and object recognition are required for feature extraction to handle the part of image processing. Hence, many relevant kinds of research on computer vision areas are derived from the feature extraction problem, such as features invariance and improving the existing methods. This article goes around the four famous issues in feature extraction and discusses solutions on the relevant research. The issues are including image feature invariance [1]*, spatial feature matching* [2]*, CNN* [3] *and R-CNN* [4] *improvement. The following explanation will compare the weakness of traditional methods and the improvement of the solutions according to the relevant researches.*

1 Introduction

Feature Extraction is an important method in Computer Vision, it is widely applied in most relevant applications, such as object recognition and image classification. It aims to extract the features from the original image and create another new feature, to reduce the number of original features.

An application such as image classification and object recognition are vulnerable to the image feature, so they emphasize the feature invariant. The image features should be invariant to the scale, rotation, illumination, etc. Additionally, the features should be highly distinctive. In other words, a single image feature should be able to match with high probability to another in a features database that contains many images accurately. By using the Scale Invariant Feature Transform to find the Scale-Invariant Keypoints of image, it can solve the two issues [1]. The implementation step and the detail explanation will be explained in Section 2.

In the image classification tasks, extracting the spatial information from an image is one of tricky issues. Usually, the recognition system applies the Bag of Features (BoF) [5] method to the computer vision tasks; however, it omits the information about the spatial layout of the image features. Spatial Pyramid Matching (SPM) is similar to BoF which also is subdividing the image repeatedly, but it computes histograms of local features at increasingly finer resolutions [2]. The detailed theory and comparison with BoF will be described in Section 3.

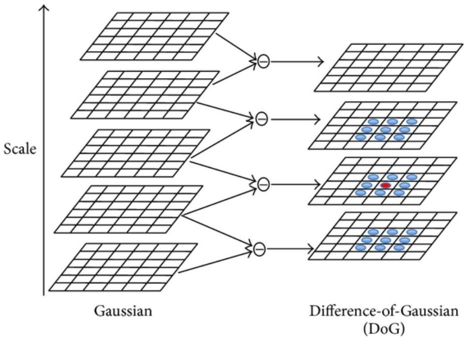
Deep Convolutional Network is widely used in the tasks of Computer Vision, such as Convolution Neural Network (CNN) which is a one of the basic models. Pooling is one of the important steps in CNN to reduce the feature map size and speed up the training. Spatial Pyramid Pooling Network (SPP-net) [3] can save more time and provide better result without content loss. Section 4 of this paper will give more details about the SSP-net.

Instance segmentation used to be a tricky task in CNN, and the existing R-CNN models cannot provide a high accuracy result. On the contrary, Mask R-CNN is the combination of Faster R-CNN and FCN on RoI [4]. It fixes the misalignment problem by RoIAlign and predicts a mask for the instance segmentation. Section 5 will discuss the details of Mask R-CNN and how it manages to improve the FR-CNN.

2 Distinctive Image Features

The image feature extraction can achieve various tasks in computer vision, such as object recognition, and image classification. Nevertheless, it requires the invariant features such as feature scaling, rotation and changing of illumination, and they should be highly distinctive, which means that the feature can be matched with the feature database correctly.

2.1 SIFT



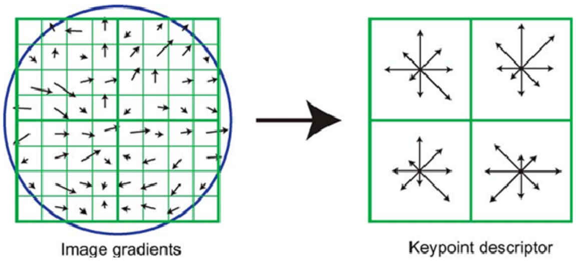


Figure 2 Keypoints Descriptor [1]

Figure 1 Difference-of-Gaussian filter [6]

Scale-Invariant Feature Transform (SIFT) is used to transform the image data to scale-invariant, it aims to find the Keypoints descriptor. Generally speaking, the Keypoints descriptor describes what feature has been changed (e.g. shape, illumination). It can be defined by the local maximum or minimum which can be discovered by using Gaussian filter to compute the difference-of-Gaussian (DoG) (see Figure 1). For getting the accurate Keypoints location, a threshold can be set to abandon the Keypoints which are unstable noise along with edges.

2.2 Scale-Invariant Keypoints

The localized Keypoints be assigned the orientation that assignment can provide the invariance for further transformation of which the computation is based on local image gradient directions. Keypoints also requires the illumination and viewpoint invariance which can be done by the Keypoints descriptor. It can be achieved by Gaussian window to weight the gradient magnitude and orientation at each pixel in a region around the Keypoints location (see Figure 2).

The Scale-Invariant Keypoints help to extract the image feature to correctly match with large database of the other Keypoints. It can produce an outstanding result to most of Computer vision task, such as object recognition [1].

3 Spatial Pyramid Matching

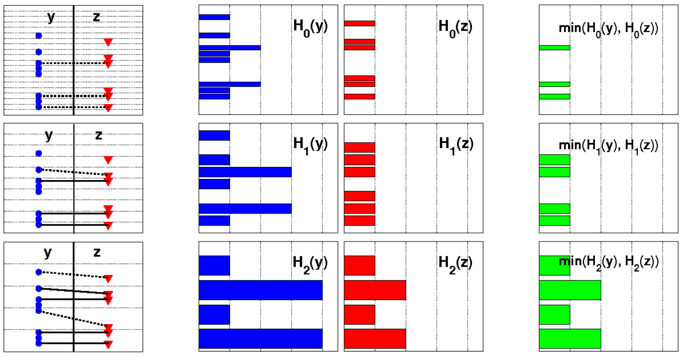
For the image classification, Bag-of-Features (BoF) is one of the common methods to be widely used. However, its result of feature extraction cannot keep the scene information. Spatial Pyramid Matching (SPM) is a remarkable solution to show extracting the features with keeping the spatial information by the spatial matching scheme which is extended from part of BoF.

3.1 Bag-of-features

Bag-of-Features (BoF) is an approach of representing an image as an orderless collection of local features. Due to its simplicity, it tends to leak any structure or spatial information of image [5]. This limitation also inhibits the BoF from capturing the shape or segment the object from image background. As for setting up an effective recognition system, it needs the ability of working in the heavy clutter, occlusion or large viewpoint changes situation [2]. It will be a hindrance in image representation task by using BoF.

3.2 Spatial Matching Scheme

As mentioned, Spatial Matching Scheme is subdividing the image repeatedly and computing the histograms of local feature at increasingly fine resolution. It is different from BoF which will disregards the spatial information of features. In contrast, the matching scheme has considered the spatial information by comparing the features intersections over different levels of resolution.



(a) Point sets (b) Histogram pyramids (c) Intersections

Figure 3 Pyramid Match in two 1-D feature sets [7]

The matching scheme extends the pyramid match kernels method (see Figure 7). The pyramid match kernels find the approximate correspondences between two sets of vectors in different dimensional features space. It places the grids over the feature space that starts from coarse to finer and take the weighted sum of matches that are found on each level of feature space. The resultative weighted sum is proved to be higher when it was found in finer resolution than it was found in coarser resolution. In pyramid matching, the finer matching is more reliable than coarse matching.

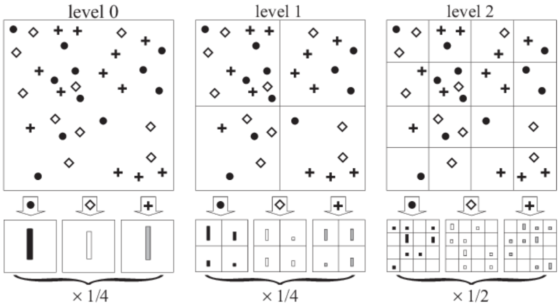


Figure 4 Example of constructing a three-level pyramid [2]

Generally speaking, the main difference between pyramid match and SPM is that SPM uses the image instead of feature space. SPM uses the pyramid match method to find the matches but execute on the image space. The matching is achieved by constructing a *M* type of visual vocabulary which is like the BoF, and it is trained in the feature space. Only the same type of features can be matched according to the trained visual vocabulary. For the feature extraction, SPM uses the dense SIFT descriptors instead of the oriented edge points to improve the sence classification. For instance, the level 2 of resolution (see Figure 4) is divided by 16 cells; each cell has a spatial bin which represents the accumulation of histogram of the matched features and then computes the weight of each spatial histogram. To sum up, SPM is a simple and efficient method for extracting the spatial layout and sence information of the image.

4 Feature Extraction in CNN

Convolutional Neural Network (CNN) is a class of DNN. However, it requires the fixed input image size which is one of the drawbacks to the accuracy of result. Spatial Pyramid Pooling Network (SPP-net) simplifies the feature extraction part in CNN to generate a fixed-size output to FCN regardless of the input image size and also improves the overall performance.

4.1 CNN

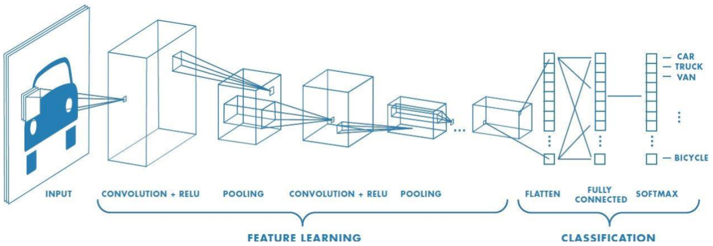


Figure 5 Architecture of CNN

Convolutional Neural Network (CNN) is a class of deep learning network. It has two parts which are feature extraction and classification. The feature extraction consists of the convolution layer with the use of ReLu activation function and the pooling repeatedly, and the classification part consists of the Fully Connected Network (FCN) (see Figure 5). The weakness of CNN is that FCN requires the fixed input image size, so before inputting the image into the CNN which is required to be cropped or wrapped to fit into the FCN. It may result in content loss and scale varying.

4.2 Spatial Pyramid Pooling

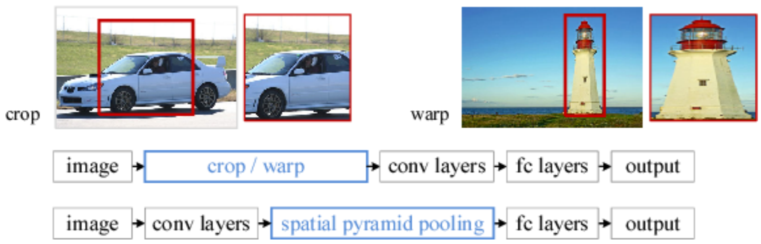


Figure 6 Top: Cropping and Wrapping the image to fit the FCN.

Middle: CNN. Bottom: SPP-net [3]

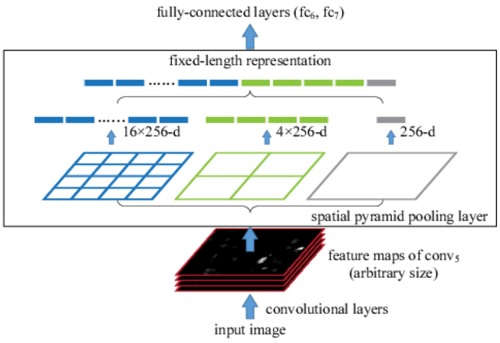

fully-connected layers (fc6, fc7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of conv5
(arbitrary size)
convolutional layers
input image

Figure 7 SPP-net Structure [3]

Spatial Pyramid Pooling Network (SPP-net) is a network which is based on the CNN; it uses the SPM (see Section 3) to remove the fixed-size constraints and the repeated convolutions. As mentioned, the fixed-size constraint in CNN occurs in the FCN layers, but convolutional layers can generate the features no matter what the input size is.

The operation of SPP-net is that only one convolution is done to the image on feature extraction and replace the last pooling layer to the SPP. SPP places the grid from coarser to finer over the feature map, and there will be different number of cells depending on the level of resolutions (see Section 3). Every cell has a spatial bin; it pools the response of each filters which is used in the convolutional layer by using max or average pooling. The outputs of SPP are the vectors of which the dimension is m numbers of bins × k filters in convolutional layers which can fit into the FCN without any content loss in any scales and ratios.

SPP-net is a remarkable and efficient method to handle the images and avoid the content loss by forcing the image size to be fitted into the FCN.

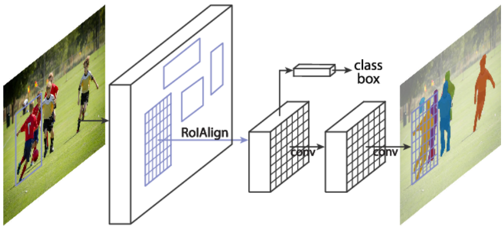
5 Instance Segmentation

The feature extraction is also important to the object detection, such as the Region CNN (R-CNN). One of traditional methods is R-CNN and its extension such as Faster R-CNN. However, these methods cannot provide an accuracy result in instance segmentation. The Mask R-CNN combines the Faster R-CNN and FCN to implement the instance segmentation tasks.

5.1 R-CNN

R-CNN is used in object detection; it selects the region proposals and then extract the features by CNN and classify the result by SVMs, but it is too slow. So, there are Fast R-CNN and Faster R-CNN which both are much faster than R-CNN. Fast R-CNN only does once CNN and uses the RoIPool to extract the features. Faster R-CNN, as its name suggests, is faster. It uses Region Proposal Network (RPN) to generate the region proposals on the images. However, they cannot perform well in instance segmentation because of their segmentation before the recognition which is slow and less accurate.

5.2 Mask R-CNN



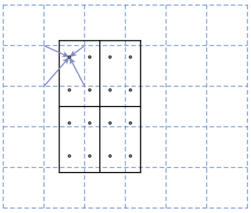


Figure 8 Mask R-CNN [4]

Figure 9 Implementation of RoIAlign [4]

Mask R-CNN is an extension of Faster R-CNN and FCN; it is based on parallel prediction of masks and class labels [4] which can remove the drawbacks being mentioned in Section 5.1. The major difference of Mask R-CNN is that it uses the RoIAlign to replace the RoIPool, and each of region will predict a mask to perform the segmentation (see Figure 8).

Each RoI in Mask R-CNN predict a binary mask by FCN. It is used in performing the segmentation in the final result to each class. The mask encodes the spatial structure of input object's spatial layout; the convolution can solve the pixel-to-pixel correspondence when the mask is performing on the result.

In Faster R-CNN, the RoIPool gives the misalignments, and it is usually not a big problem in the detection. But the instance segmentation requires pixel-to-pixel alignment while the predicting mask requires the finer spatial localization. RoIAlign can align the extracted features with the input. In figure 9, RoIAlign uses bilinear interpolation to compute the exact values of the input features at four regularly sampled locations in each bin and aggregates the result by max or average pooling.

The RoIAlign improves the result accuracy, and the predicted mask can achieve the instance segmentation by using the Faster R-CNN.

6 Conclusion

Feature extraction is important in computer vision no matter what task it is. Most of computer vision tasks emphasize the feature invariance, and it can be achieved by using SIFT to find the Keypoints descriptors. Most of relevant researches focus on improving its performance in different tasks, such as applying the Spatial Pyramid Pooling in CNN to reduce the content loss. Simultaneously, the Spatial Pyramid Matching can provide the spatial information of the features. Some of popular models can achieve another purpose through changing its feature extraction part, such as Mask R-CNN which changes the pooling layer and adds a branch to predict a mask to be used in instance segmentation.

References:

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision 60,* p. 91–110, 2004.

[2] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for recognizing natural scene categories," *CVPR'06,* pp. 2169-2178, 2006.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in deep convolutional networks for visual recognition," *ECCV,* pp. 346-361, 2014.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV),* pp. 2980-2988, 2017.

[5] S. O'Hara and B.A. Draper,, *Introduction to the Bag of Features paradigm for image classification and retrieval,* arXiv:1101.3354 [cs.CV], 2011.

[6] M. Huang, Z. Mu, H. Zeng, and H. Huang, "A novel approach for interest point detection via Laplacian-of-Bilateral Filter.," *Journal of Sensors,* pp. 1-9, 2015.

[7] K. Grauman and T. Darrell, "Pyramid Match Kernels: Discriminative classification with sets of image features," *ICCV,* 2005.