Computer Vision Homework 3

Implement Bag of visual words and Spatial Pyramid Matching method

Wencai Jiang
Shanghaitech University
jiangcw@shanghaitech.edu.cn

Abstract

This report introduces how to implement the algorithm for Recognizing Natural Scene Categories by Bag of visual words[3] and Spatial Pyramid Matching [2]. I test algorithm on the publicly available Caltech-256[1].

1. Introduction

1.1. Bag-of-Visual-Words

Using image keypoints or local interest points in image retrieval and classification. Keypoints are salient image patches that contain rich local information of an image, and they can be automatically detected using various detectors and represented by many descriptors [. Keypoints are then grouped into a large number of clusters so that those with similar descriptors are assigned into the same cluster. By treating each cluster as a "visual word" that represents the specific local pattern shared by the keypoints in that clus- ter, we have a visual-word vocabulary describing all kinds of local image patterns. With its keypoints mapped into vi- sual words, an image can be represented as a "bag of visual words", or specifically, as a vector containing the (weighted) count of each visual word in that image, which can be used as a feature vector in classification task.

1.2. Spatial Pyramid Matching

Although the Bags of Features has great performance, it is weak in spatial information. People can construct some strange images that consist of some of features without caring about the spatial structure, which can also be clas- sified as a targeted label with high probabil- ity. In this paper, the author proposed a new method that contain the spatial information, which called Spatial Pyramid Matching(SPM).

2. Program analysis

In this section we analyze the program step by step, and discuss the related parameters setting.

2.1. Bag-of-Visual-Words

we use the vector quantization (VQ) technique which clusters the keypoint descriptors in their feature space into a large number of clusters using the K-means clustering algorithm and encodes each keypoint by the index of the cluster to which it belongs. We conceive each cluster as a visual word that represents a specific lo- cal pattern shared by the keypoints in that cluster. Thus, the clustering process generates a visual-word vocabulary de- scribing different local patterns in images. The number of clusters determines the size of the vocabulary, which can vary from hundreds to over tens of thousands. By mapping the keypoints to visual words, we can represent each image as a "bag of visual words". This representation is analo- gous to the bag-of-words document representation in terms of form and semantics. Both representations are sparse and high-dimensional, and just as words convey meanings of a document, visual words reveal local patterns characteristic of the whole image.

2.2. Spatial Pyramid Matching

First,we need to extract keypoints by SIFT and build codebook by KMean.For each RBG image, get gray image firstly and calculate the key points for $stepsize$: 4. Then, calculate the descriptors by key points. For all images, I got $X_train_feature$.

Set $K = 100$, use K-means to build codebook with $X_train_feature$.

Then, Building Spatial Pyramid.Construct a sequence of grids at resolutions $l = 0, 1, 2$, such that the grid at level l has 2^l cells along each dimension, for a total of $D = 2^{dl}$ cells. Finally, for each level,

concatenate the pyramid, and calculate the histogram with a weight.

3. Result

3.1. Spectral Residual Approach

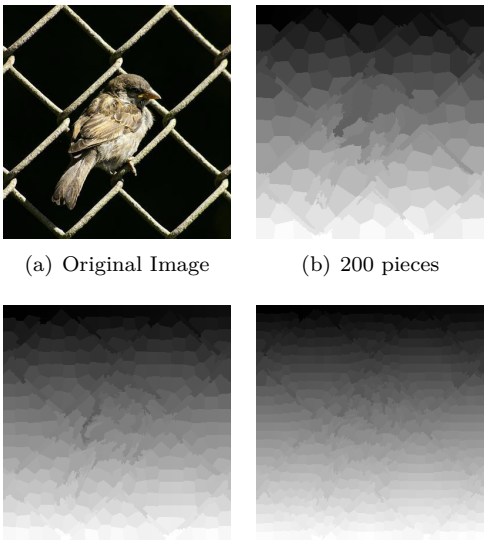
Here are the results of spectral residual approach shown in figure 1

(a) Original Image (b) Saliency map

Figure 1. Left: input; Right: the result

3.2. Context-Aware Saliency Detection

First, Using the existing code on the Internet, I try to divide the picture into different number of super pixels the image segmentation results shown in Figure 2.



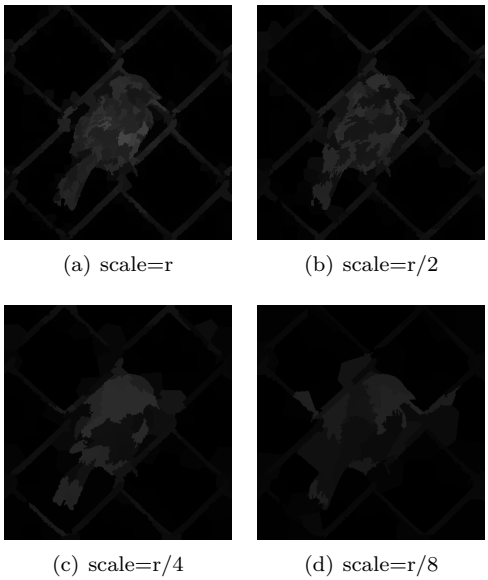
(a) Original Image (b) 200 pieces

(c) 400 pieces (d) 800 pieces

Figure 2. Image split into 200 400 and 800 pieces

Then I use 4 scales in the program to calculate the saliency:$R = r, \frac{1}{2}r, \frac{1}{4}r, \frac{1}{8}r$. Saliency of patches under different scales are shown in Figure 3.

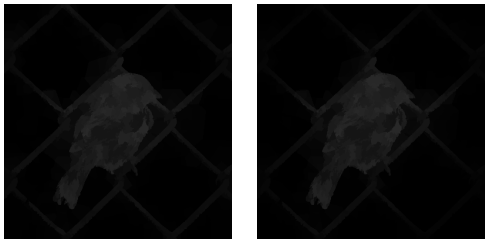
Finally Taking the mean of the saliency at pixel i at different scales and giving pixel saliency weight according to their distance to the foci to get the final computed saliency . The results shown in Figure 4.



(a) scale=r (b) scale=r/2

(c) scale=r/4 (d) scale=r/8

Figure 3. Saliency of patches under different scales



(a) mean pixel (b) weighted pixe

Figure 4. Taking the mean of the saliency at pixel i and giving pixel saliency weight

References

[1] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[2] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In Computer vision and pattern recognition, 2006 IEEE computer society conference on, volume 2, pages 2169–2178. IEEE, 2006.

[3] J. YANG. Evaluating bag-of-visual-words representations in scene classification. Proc Mir, 2007.