Computer Vision

Project Report

“Scene Recognition using Spatial Pyramid Matching over Bag of Words Model”

**Team Members:**

1. Ayman Mohamed
2. AbdelRahman Ahmed
3. AbdelRahman Fawzy
4. AbdelRahman Nasr

**Introduction**

In this report we discuss the methodology and results we worked on to solve the problem if recognizing the semantic category of an image as described in the project description, using Spatial Pyramid matching adapted from pyramid matching scheme of Grauman and Darrell [1] over basic bag of words representation [2], in the presence of occlusion, shadows and cluttering. The system is implemented in C++ program on Linux system using OpenCV library for computer vision operations.

**Project Description**

In this section we illustrate the project specifications required and the algorithm description and how it is implemented in C++ OpenCV on Linux based system. The Goal of the project is to identify the scene in the input image from 15 different scene types, using C++ OpenCV library on Linux based system, the algorithm is based on bag of words model [2] with an extension to encode part of the spatial information in the histogram of the visual words [3] (Bag).

The dataset is the scenes of different 15 scene categories, 13 were provided by Fei-Fei and Perona [4] (8 of these were originally collected by Oliva and Torralba [5]), and 2 where collected by Lazebnik, Schmid and Ponce [2]. Each Category has 200 to 400 grayscale images, and average image size is 300 x 250 pixels. The classification experiment is done using 100 images per-class for training and the rest for testing (the same setup as [2] and [4]).



Figure 1. Example image from the scene category database.

The algorithm is implemented as shown in the figure, the Training Set is first processed to get the features of each image using one of the Feature Extractor Algorithms (SIFT [6], SURF [7], ORB), then the outputted features are clustered using k-mean clustering algorithm [8] to get the centers of each cluster to be used as a visual word in the bag of words. This process is the initialization of the whole algorithm parameters.

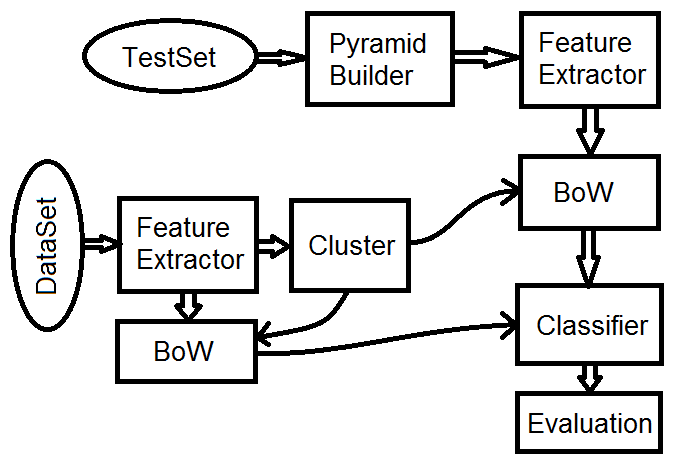


Figure 2. The Algorithm’s program flow flowchart.

The random Training set and the Test set is processed using a Pyramid builder algorithm [1] to get the spatial blocks of each image to be processed, the Feature extractor is executed on the spatial blocks of each pyramid level to get the final features of the image.

The new features are classified according to its distance from the visual words of the bag of words in the BoW algorithm that depends on the geometric distance between each feature and the list of visual words, to get the outputted representation of the whole image as histogram of the Bag of visual words [9], including some details of the spatial information of the image.

The Classifier algorithm (SVM [10]) is then used to classify the histogram [9] of the image to get its predicted scene according to histograms of the Training set that trains the classifier algorithm.

**Experiments**

In this section we report our experiment results over the dataset images repeated 5 times with different randomly selected training and test images, and the average of per-class recognition rates is recorded in each run. The final result is reported as geometric mean and standard deviation of the results from the individual runs.

Using 100 clusters for the bag of words size with SURF feature extractor and linear SVM Classifier, the performance is increased significantly with the increase of the number of pyramid levels, as the average accuracy of the Training set classification mean increases from 86% to 100% with the increase of the number of levels from 0 to 1 level. While the Test set mean increases from 48.7% to 55.58% with the increase of the number of levels from 1 to 2.

The SIFT feature extractor outputs doesn’t differ from SURF feature extractor as the SIFT outs 49.3% for 2 level pyramid and 100 cluster. While the ORB feature extractor performs less performance about 15% for 1 level pyramid and about 30% for 3 level pyramid.

Increasing the size of the bag of words histogram (number of clusters k) doesn’t improve too much as in 2 levels pyramid and linear SVM the mean Test set outs 58.26%

The changing in SVM classifier scheme changes in the results as changing from Linear SVM to Polynomial SVM increases the accuracy slightly but the computations are increased too.

**Results**

In the End the most efficient configuration of the system is using the k-means clustering algorithm with k equals 200 on the SURF feature extractor, on the SVM classifier with Linear SVM classification that results mean percentage of 58.26% and standard deviation of 0.5% for 5 random runs.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 57 | 0 | 16 | 16 | 25 | 5 | 0 | 0 | 2 | 3 | 1 | 0 | 10 | 11 | 4 |
| 1 | 100 | 4 | 1 | 3 | 0 | 0 | 2 | 3 | 2 | 4 | 2 | 2 | 0 | 1 |
| 14 | 10 | 64 | 6 | 4 | 11 | 0 | 5 | 15 | 13 | 14 | 9 | 34 | 1 | 11 |
| 13 | 5 | 10 | 48 | 18 | 1 | 0 | 1 | 21 | 1 | 0 | 2 | 0 | 21 | 8 |
| 10 | 9 | 12 | 8 | 70 | 0 | 0 | 1 | 5 | 2 | 0 | 5 | 8 | 10 | 14 |
| 2 | 0 | 5 | 0 | 0 | 180 | 0 | 8 | 1 | 13 | 43 | 1 | 9 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 2 | 209 | 0 | 0 | 6 | 11 | 1 | 1 | 0 | 6 |
| 0 | 0 | 2 | 2 | 0 | 18 | 0 | 123 | 1 | 4 | 11 | 3 | 1 | 1 | 0 |
| 3 | 4 | 30 | 7 | 17 | 1 | 0 | 3 | 95 | 3 | 1 | 14 | 21 | 4 | 21 |
| 0 | 0 | 5 | 1 | 1 | 15 | 8 | 5 | 0 | 157 | 43 | 9 | 12 | 0 | 2 |
| 2 | 4 | 7 | 0 | 0 | 25 | 6 | 2 | 4 | 40 | 172 | 2 | 8 | 0 | 1 |
| 0 | 1 | 4 | 0 | 4 | 0 | 1 | 4 | 15 | 7 | 6 | 129 | 0 | 0 | 5 |
| 1 | 2 | 11 | 0 | 4 | 1 | 0 | 2 | 8 | 5 | 2 | 2 | 134 | 0 | 1 |
| 10 | 0 | 11 | 13 | 27 | 1 | 0 | 0 | 3 | 1 | 0 | 1 | 4 | 63 | 3 |
| 3 | 5 | 30 | 8 | 16 | 0 | 4 | 4 | 35 | 17 | 2 | 12 | 12 | 4 | 138 |

Table 1. Example of Confusion Table results.

**Conclusion**

In this Report we have presented that the spatial information of the image has increased the accuracy of the whole classifier in multi-label classification. Advantages of this approach include simplicity and effective use of limited training data.

In the experiment, the data is sparse for some combined classes. We would like to apply the classification system to perform on large amount of data for each class and multiple classes. The system is demonstrated on the k-means clustering with SURF feature extraction and the SVM classifier, but we are interesting in generalizing our methods to other classifiers and feature extractors.

**References**

[1] K. Grauman and T. Darrell. Pyramid match kernels: Discriminative classification with sets of image features. In *Proc. ICCV*, 2005.

[2] Tao Mei, In-So Kweon, Xian-Sheng Hua. Contextual Bag-of-Words for Visual Categorization. IEEE Trans., 2011.

[3] S. Lazebnik, C. Schmid and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories.

[4] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, 2005.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[6] D. Lowe. Towards a computational model for object recognition in IT cortex. In *Biologically Motivated Computer Vision*, pages 20–31, 2000.

[7] Su Juan, Xu Qingsong, Zhu Jinghua. A scene matching algorithm based on SURF feature. IASP, 2010.

[8] H. Zhang, A. Berg, M. Maire, and J. Malik. SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *Proc. CVPR*, 2006.

[9] E. Hadjidemetriou, M. Grossberg, and S. Nayar. Multiresolution histograms and their use in recognition. *IEEE Trans.* *PAMI*, 26(7):831–847, 2004.

[10] M. R. Boutell, J. Luo, Xipeng Shen, Christopher M. Brown. Learning multi-label scene classification. 2004.