



國立臺灣科技大學  
資訊工程系

---

## 碩 士 論 文

A Learning-based Algorithm for Natural Scene Recognition

研 究 生：Dang Duy Thang

學 號：M10215804

指導教授：花凱龍博士

中華民國一百零四年七月三十日

## Abstract

Scene recognition is an important problem in many application areas of image and video processing. Scene recognition has a wide range of applications, such as object recognition and detection, content-based image indexing and retrieval and intelligent vehicle and robot navigation. However, the natural scene images tend to be very complex and difficult to analyze due to changes of illumination and transformation. In this thesis, we will investigate into building a novel model to learn and recognize scenes in nature.

This study proposes a new approach that combines locality-constrained sparse coding (LCSP), Spatial Pyramid Pooling and linear SVM in end-to-end model. Firstly, interesting points of each image in the training set are extracted by a local descriptor as dense SIFT which represents local spatial information. These features known as codewords and each codeword is represented as part of a topic. Then we employ LCSP algorithm to learn the codeword distribution of those local features from the training dataset. Next, a modified Spatial Pyramid Pooling model is employed for encoding the spatial distribution of local features. Spatial Pyramid Pooling model has been remarkably successful in terms of both scene and object recognition. In the testing stage, a linear SVM will be used to classify local features which are encoded by Spatial Pyramid Pooling. The new system achieves very competitive results and leads to state-of-the-art performance on several benchmarks.

# Acknowledgements

Studying master of computer science at National Taiwan University of Science and Technology was one of the best decisions of my life. I would like to thank Prof. Kai-Lung Hua for the enthusiastic support and for many things that he has taught me. His kindness to meet at any time of day was crucial to my researching process and coming up with new advice.

Additionally, this work would not have been possible without the help of all the students at Video and Image Processing Laboratory (VIP lab) for the interesting discussions and research they have all done.

I would also like to thank my family for their continuing support in my studies and helping me with all the tough decisions along the way. I thank my parents, Dang Duy Quyet, Tran Thi Kim Thanh and my both of elder sisters, Thu Trang and Thanh Hoa who always encouraged me and gave me many helpful comments in my life, and also thank to my wife, Quynh Trang whose patience with my long time of work.

# Table of Contents

Abstract . . . . .	i
Acknowledgements . . . . .	ii
Table of Contents . . . . .	iii
List of Tables . . . . .	v
List of Illustrations . . . . .	vi
1 Introduction . . . . .	1
1.1 Introduction of recognition of natural scene images . . . . .	1
1.2 Literature review . . . . .	2
2 Proposed Method . . . . .	4
2.1 The algorithm for recognizing natural images . . . . .	4
2.2 Extract Local Features . . . . .	6
2.3 Locality-constrained sparse coding (LCSP) . . . . .	9
2.4 Spatial Pyramid Pooling . . . . .	11
2.5 Support Vector Machine (SVM) . . . . .	13
3 Experimental Results . . . . .	16
3.1 The 8-class sports event . . . . .	17
3.2 13-category scenes . . . . .	19
3.3 15-category scenes . . . . .	20
3.4 8-class outdoor category scenes . . . . .	22
3.5 Caltech 101 dataset . . . . .	24
3.6 MIT 67-class indoor scene . . . . .	28

4	Conclusion . . . . .	32
	References . . . . .	33

# List of Tables

3.1	Accuracy rate [%] on 8-class sport event dataset . . . . .	18
3.2	Accuracy rate [%] on 13-class event dataset . . . . .	20
3.3	Accuracy rate [%] on 15-class event dataset . . . . .	22
3.4	Accuracy rate [%] on 8-class outdoor event dataset . . . . .	23
3.5	Accuracy rate [%] on Caltech 101 dataset . . . . .	28
3.6	Accuracy rate [%] on MIT 67-class indoor dataset . . . . .	31

# List of Illustrations

2.1	Given an input image, the system first extracts dense local descriptor. Then, each local descriptor is coded using LCSP. The codes of the descriptors are then passed to pooling with SPP to form feature vectors for representing the image. Linear SVM is used for forming a classification model. . . . .	6
2.2	The image shows how a SIFT point is described using a histogram of gradient magnitude and direction around the feature point . . . . .	9
2.3	The figure shows the result in feature vectors with 128 dimensions that are extracted by dense-SIFT . . . . .	9
2.4	An input vector $X$ is represented by Dictionary $D$ and sparse coefficient $c$ . . . . .	10
2.5	we subdivide the image at three different levels. The example of constructing a three-level pyramid. The image has three feature types, indicated by circles, diamonds, and crosses. Next, for each level, we calculate the features that fall in each spatial bin. . . . .	12
3.1	Local feature descriptors using: dense-SIFT, dense-COLOR and MultiBlock-LBP . . . . .	16
3.2	Images from eight different sports event categories . . . . .	17
(a)	badminton . . . . .	17
(b)	bocce . . . . .	17
(c)	croquet . . . . .	17
(d)	polo . . . . .	17
(e)	rock climbing . . . . .	17

(f)	rowing . . . . .	17
(g)	sailing . . . . .	17
(h)	snow boarding . . . . .	17
3.3	Confusion matrix of the 8-class sport event dataset . . . . .	18
3.4	Confusion matrix of the 13-class event dataset . . . . .	19
3.5	Confusion matrix of the 15-class event dataset . . . . .	21
3.6	Images from eight different outdoor categories . . . . .	22
(a)	coast . . . . .	22
(b)	forest . . . . .	22
(c)	highway . . . . .	22
(d)	insidecity . . . . .	22
(e)	mountain . . . . .	22
(f)	opencountry . . . . .	22
(g)	street . . . . .	22
(h)	tallbuilding . . . . .	22
3.7	Confusion matrix of the 15-class outdoor event dataset . . . . .	23
3.8	Confusion matrix of the Caltech 101 dataset . . . . .	24
3.9	Confusion matrix of the MIT 67 class indoor dataset . . . . .	29