



**Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University**

Developing a Method for 3D Scene Understanding Using Image Sequence

Thesis submitted as a partial fulfillment of the requirements
for the degree of Master of Science in Computer and
Information Sciences

By

Islam Ibrahim Fouad Ahmed

Teaching Assistant at Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Mostafa Gadai-Haqq M. Mostafa

Professor in Computer Science Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Sherine Rady Abdel Ghany

Assistant Professor in Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

2017

Acknowledgment

I acknowledge my deep gratitude to ALLAH for providing me with the strength to complete this work on a level that I hope will please the reader.

I wish to express my sincere and heartfelt appreciation to my supervisor; Prof. Mostafa Gadal-Haqq for his supervision, resourceful ideas and suggestions that helped in enhancing my work.

Special thanks are due to my supervisor Dr. Sherine Rady for her continuous guidance and encouragement in achieving this work.

Without their assistance and dedicated involvement in every step throughout the process, this thesis would have never been accomplished. I would like to thank them very much for their support and understanding.

Finally, with all my appreciation and love that no words can describe, my deepest gratitude goes to my family for their love, prayers and endless support.

Abstract

Indoor scene understanding is a challenging problem in computer vision. To achieve an accurate solution for this task, a model that can exploit discriminating information between different scene categories and objects is necessary.

This thesis presents a framework for scene understanding which includes several components of learning models, segmentation, object recognition and tracking. A comprehensive study for supervised learning models for recognizing indoor scenes is presented. The study compares between several “Shallow Learning” models against the recent approach “Deep Learning”. Furthermore, the robustness of methods is tested against environment changes such as: contrast degradation, additive blurring and additive noise.

A segmentation method is proposed for object recognition which relies on depth for accurate segmentation preprocessing before applying learning models. The process is applied on both RGB and depth images to produce two segmented images. Finally, the result of both segmented images is combined.

For object recognition, a hierarchical object recognition scheme is proposed based on multiple instances

of shallow learning models. The first level in hierarchy classifies objects based on the scene category, while the second level classifies the objects based on their occupied area, and the third and last level classifies the objects based on Histogram of Oriented Gradient descriptor of each object.

The recognized objects are tracked in image sequence by extracting Shi Tomasi's good features, then the new location for these extracted features are to be located in the next image using pyramid Lucas and Kanade tracker. Finally, a method for defining the relation between recognized objects in the scene is proposed.

All the proposed methods in the framework have been tested on the standard benchmark MIT-indoor datasets. Four experiments are presented: the first experiment compares Shallow vs Deep Learning. Experiments shows that Deep learning models outperform the shallow learning models by a huge gap with respect to classification accuracy, especially the deep architecture called "VGG-16 net" outperforms all techniques. The additional step of including a fine-tuning for a pre-trained "VGG-16 net" proved to be highly significant in improving the classification accuracy that reached 85.43%, and at a cost of 20% reduction savings in training time.

The second experiment tests the robustness of presented learning models. The HOG-SVM shows outstanding robustness against contrast degradation and blurring more than deep learning and less time and space, despite the lower accuracy performance.

The third experiment tests the proposed segmentation algorithm. Results show effectiveness with respect to average execution time that reaches 11sec per image, which is faster than other algorithms in related work. Experimentation with the RGB-D dataset consisting of 1200 images “NYU V2” of indoor scenes showed that 73.41% are correctly segmented images.

The Fourth experiment shows that the proposed hierarchical classification reaches 65.73% classification accuracy. The result of this hierarchical classification is used for training each scene category in the multiple deep learning model VGG-16 net. The achieved accuracy for scene recognition is 89.24%. Performing Tracking over an image sequence for a sequence of 10 frames proved to improve the speed performance of the indoor scene understanding framework by ten times, with the process applied on first frame only and then only on the new part of scene including the new objects.

Keywords: Scene understanding, image recognition, supervised learning, shallow learning, deep learning, RGB-D image, image segmentation, object recognition, tracking.

List of Publications

- 1- Fouad, I. I., Rady, S., & Mostafa, M. G., “Indoor Scene Classification: A Comparative Study of Feature Detectors and Local Descriptors,” *Proceedings of the 10th International Conference on Informatics and Systems*, 2016, pp. 215-221.
- 2- Fouad, I. I., Rady, S., & Mostafa, M. G., “Enhancing RGB-D Image Segmentation using Edge Detection and Morphological Operations” *Proceedings of the 12th International Conference on Computer Engineering and Systems ICCES*, 2017, pp. 353-358.

Table of Contents

Acknowledgment	II
Abstract.....	III
List of Publications	VII
Table of Contents	VIII
List of Figures	XII
List of Tables	XVIII
List of Abbreviations	XVIII
Chapter 1. Introduction	2
1.1 Overview	2
1.2 Motivation	5
1.3 Objectives.....	6
1.4 Methodology	7
1.5 Contributions.....	9
1.6 Thesis Organization	10
Chapter 2. Background and Related work.....	14
2.1 Background	14
2.1.1 Image Recognition Systems.....	14
2.1.2 Learning Models	15
2.1.3 RGB-D Images.....	17
2.1.4 Image Segmentation.....	18
2.1.5 Object Recognition	19
2.2 Related Work.....	19
2.2.1 Vision-based Scene recognition.....	20
2.2.1.1 Scene recognition using low level features.....	20
2.2.1.2 Combined local & global features for scene recognition	25
2.2.1.3 Scene recognition through object detection.....	26
2.2.1.4 Scene recognition using deep learning	27
2.2.2 RGB-D indoor scene segmentation ...	28

Chapter 3.	The study of Shallow Learning techniques for indoor scene classification	32
3.1	Feature Extraction	32
3.1.1	SIFT Feature Detector/Descriptor	33
3.1.2	SURF Feature Detector/Descriptor....	37
3.1.3	FAST Feature Detector	41
3.1.4	BRISK Feature Detector/Descriptor..	43
3.1.5	ORB Feature Detector/Descriptor	46
3.1.6	MSER Feature Detector	48
3.1.7	HOG Local Descriptor	49
3.1.8	LBP Local Descriptor	50
3.2	Classification.....	52
3.2.1	SVM.....	52
3.2.2	K-NN.....	55
3.3	Experiments and results	56
3.3.1	Performance analysis of different feature extraction / classification methods.....	59
3.3.2	Effect of image distortions caused by environment	64
3.3.2.1	Contrast Degradation	64
3.3.2.2	Additive Blurring.....	67
3.3.2.3	Additive Gaussian Noise	69
3.4	Summary and Conclusion	71
Chapter 4.	Deep Learning for indoor scene recognition.	74
4.1	Deep Learning Architectures	77
4.1.1	GoogLeNet.....	77
4.1.2	VGG-16 Net.....	81
4.1.3	PlacesCNN	85
4.2	Experiments and results	86
4.2.1	Performance of deep learning vs. shallow learning.....	86
4.2.2	Performance of pre-trained models ...	89
4.2.3	Fine-tuning CNN models.....	92
4.2.4	Effect of image distortions caused by environment	93

	4.2.4.1	Contrast degradation.....	94
	4.2.4.2	Additive Blurring.....	96
	4.2.4.3	Additive Gaussian noise	98
4.3		Summary and Conclusion	99
Chapter 5.		Segmentating 3D scene and classifying objects in image sequence	104
5.1		RGB-D Indoor Scene Dataset	104
5.2		Indoor 3-D Scene Segmentation	106
	5.2.1	Depth Image Segmentation.....	107
	5.2.1.1	Region Growing method..	107
	5.2.1.2	Segmentation using Object boundaries	109
	5.2.1.3	Segmentation using Object boundaries improvement..	110
	5.2.2	RGB Image Segmentation	115
	5.2.3	RGB-D Image Segmentation	116
5.3		Indoor Scene Object Classification.....	119
	5.3.1	Scene objects direct classification ...	121
	5.3.2	Hierarchical scene object classification	125
	5.3.3	Hierarchical scene object classification with constraints.....	125
	5.3.4	Object classification using deep learning	126
5.4		Recognized objects tracking in image sequence	127
5.5		Summarizing image sequence in a semantic description	131
5.6		Summary and Conclusion	134
Chapter 6.		Conclusion and Future Work	137
6.1		Conclusion.....	137
6.2		Future work	141
References.			143

List of Figures

Fig. 1.1	Illustration of scene understanding process.....	9
Fig.2.1	Sample RGB and depth images taken by Microsoft Kinect.....	18
Fig. 2.2	Color-based feature for (a) city and (b) for landscape image; (c) & (d) show the color histogram features for (a) & (b); (e) & (f) show the coherent color bins for (a) & (b); (g) & (h) show the non-coherent color bins for (a) & (b).	21
Fig. 2.3	Edge-based feature for (a) city and (b) for landscape image; (c) & (d) show the edge direction histogram features for (a) & (b); (e) & (f) show the coherent edge direction bins for (a) & (b); (g) & (h) show the non-coherent edge direction bins for (a) & (b).	22
Fig. 2.4	2D Feature space plots showing (a) edge detection coherence vector and (b) coherence vectors; * represents the landscape patterns and $\diamond$ represents the city pattern; only a subset of 2716 patterns have been plotted here for clarity of display.	23
Fig. 2.5	Two-stage classification combining color and texture, such that "in" means indoor and "out" means outdoor [30].	24
Fig. 2.6	Example of a scene prototype. a) Scene prototype with candidate ROI. b) Illustration of the visual words and the regions used to compute histograms.	25
Fig. 2.7	Multiclass average precision performance for the baseline and four different model versions.....	26
Fig .2.8	The results of scene recognition through object detection on an office environment.....	27

Fig. 2.9	Joint deep multiple instance learning framework for learning correspondences between keywords and image regions	28
Fig. 2.10	Illustration of proposed framework for jointly learning image regions and keywords. Here P stands for a pooling layer, C for a convolution layer, and F for a fully connected layer.	28
Fig. 2.11	Illustration of RGB-D segmentation.....	30
Fig. 2.12	Sample output for the proposed method [45]. (a) RGB image, (b) Depth image, (c) Result of segmentation.....	30
Fig. 3.1	Difference of Gaussian pyramid.....	34
Fig. 3.2	Maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel (marked with X) to its 26 neighbors.	35
Fig. 3.3	SIFT keypoint descriptor.	37
Fig. 3.4	Using integral images, it takes only three additions to calculate the sum of intensities inside a rectangular region of any size.....	38
Fig. 3.5	Iteratively reducing the image size (left). The use of integral images allows the up-scaling of the filter at constant cost (right).	39
Fig. 3.6	Haar wavelet filters to compute the responses in x (left) and y direction (right). The dark parts have the weight -1 and the light parts +1.....	40
Fig. 3.7	Orientation assignment in SURF.	40
Fig. 3.8	To build the descriptor, an oriented quadratic grid with 4×4 square sub-regions is laid over the interest point (left). For each square, the wavelet responses are computed.	41

Fig. 3.9	Image showing the interest point under test and the 16 pixels on the circle	42
Fig. 3.10	Scale-space interest point detection in BRISK.....	43
Fig. 3.11	The BRISK sampling pattern	44
Fig. 3.12	Circularly symmetric neighbor sets. Samples that do not exactly match the pixel grid are obtained via interpolation.....	50
Fig. 3.13	Different texture primitives detected by the LBP.....	51
Fig. 3.14	SVM finds the optimal hyperplane that maximize the margin between all samples in training set.	53
Fig. 3.15	Red dot is a testing sample is tested with different values for k, such that '+' is a category and '-' is another category ..	56
Fig. 3.16	Sample images of CVPR_09 dataset.....	57
Fig. 3.17	Accuracy of BRISK, ORB, SIFT and SURF techniques with respect to changes in image contrast degradation percentage.....	66
Fig. 3.18	Accuracy of FAST-BRIEF, FAST-BRISK, FAST-ORB and LBP techniques with respect to changes in image contrast degradation percentage.....	66
Fig. 3.19	Accuracy of MSER-BRIEF, MSER-BRISK, MSER-ORB and HOG techniques with respect to changes in image contrast degradation percentage.	67
Fig. 3.20	Accuracy of BRISK, ORB, SIFT and SURF techniques with respect to changes in Gaussian blurring sigma.	68
Fig. 3.21	Accuracy of FAST-BRIEF, FAST-BRISK, FAST-ORB and LBP techniques with respect to changes in Gaussian blurring sigma.	68

Fig. 3.22 Accuracy of MSER-BRIEF, MSER-BRISK, MSER-ORB and HOG techniques with respect to changes in Gaussian blurring sigma.....	69
Fig. 3.23 Accuracy of BRISK, ORB, SIFT and SURF techniques with respect to changes in noise ratio.....	70
Fig. 3.24 Accuracy of FAST-BRIEF, FAST-BRISK, FAST-ORB and LBP techniques with respect to changes in noise ratio.	70
Fig. 3.25 Accuracy of MSER-BRIEF, MSER-BRISK, MSER-ORB and HOG techniques with respect to changes in noise ratio.....	71
Fig. 4.1 General architecture for Convolutional Neural Network ..	76
Fig. 4.2 GoogLeNet network with all layers.....	78
Fig. 4.3 Inception module	79
Fig. 4.4 PlacesCNN Architecture.	85
Fig. 4.5 Accuracy of all techniques with respect to changes in image contrast degradation percentage.	95
Fig. 4.6 Accuracy of all techniques with respect to changes in Gaussian blurring sigma.	97
Fig. 4.7 Accuracy of all techniques with respect to changes in noise ratio.....	99
Fig. 4.8 Example of transformed images. For each image, we also show the output of the soft-max unit for the correct class. This output corresponds to the confidence the network has of the considered class. For all networks and for all transformations this confidence decreases as the image quality decreases.....	101
Fig. 5.1 Sample RGB images (L.H.S) and Depth images (R.H.S) for kitchen scene	105

Fig. 5.2	Sample RGB images (L.H.S) and Depth images (R.H.S) for bedroom scene.....	106
Fig. 5.3	Sample output of region growing algorithm with different values for T and elimination size.....	108
Fig. 5.4	(a) Depth image, (b) Sobel edge magnitude, (c) Negative of (b) and (d) Result of connected component on image (c).....	109
Fig. 5.5	Result of dilation followed by erosion with different SE size.....	110
Fig. 5.6	Result of segmentation without dilation and erosion on the left and with on the right	111
Fig. 5.7	(a) Erosion and dilation of Sobel edge magnitude, (b) Erosion of (a) by SE = 25, (c) Connected component of (b) with elimination of small objects	112
Fig. 5.8	(a) Result from Figure 5.6, (b) Result from Figure 5.7, (c) Combining components of both (a) & (b), finally (d) is the result of merging components that is related to each other based on the spatial and depth criterion	114
Fig. 5.9	(a) RGB image, (b) Result of RGB image segmentation	115
Fig. 5.10	(a) RGB image, (b) Depth image, (c) Segmentation result of RGB image, (d) Segmentation result of Depth image and (e) is the result of combining both RGB and Depth image segmentation.....	117
Fig. 5.11	Illustration of the proposed RGB-D segmentation. Blocks in the same row are running on different threads as they are independent on each other. Blocks in same column are applied sequential	118