UNIVERSITY OF TECHNOLOGY SYDNEY

DOCTORAL THESIS

Feature Fusion, Feature Selection and Local N-ary Patterns for Object Recognition and Image Classification

Author:
Sheng WANG

Supervisor:
Qiang WU, Xiangjian HE

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

in the

School of Computing and Communications
Faculty of Engineering and Information Technology

January 2015

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Student: ____________________

Date: ____________________

“Concern for man and his fate must always form the chief interest of all technical endeavors. Never forget this in the midst of your diagrams and equations.”

Albert Einstein

UNIVERSITY OF TECHNOLOGY, SYDNEY

Abstract

Faculty of Engineering and Information Technology

School of Computing and Communications

Doctor of Philosophy

Feature Fusion, Feature Selection and Local N-ary Patterns for Object Recognition and Image Classification

by Sheng WANG

Object recognition is one of the most fundamental topics in computer vision. During past years, it has been the interest for both academies working in computer science and professionals working in the information technology (IT) industry. The popularity of object recognition has been proven by its motivation of sophisticated theories in science and wide spread applications in the industry. Nowadays, with more powerful machine learning tools (both hardware and software) and the huge amount of information (data) readily available, higher expectations are imposed on object recognition. At its early stage in the 1990s, the task of object recognition can be as simple as to differentiate between object of interest and non-object of interest from a single still image. Currently, the task of object recognition may as well includes the segmentation and labeling of different image regions (i.e., to assign each segmented image region a meaningful label based on objects appear in those regions), and then using computer programs to infer the scene of the overall image based on those segmented regions. The original two-class classification problem is now getting more complex as it now evolves toward a multi-class classification problem. In this thesis, contributions on object recognition are made in two aspects. These are, improvements using feature fusion and improvements using feature selection. Three examples are given in this thesis to illustrate three different feature fusion methods, the descriptor concatenation (the low-level fusion), the confidence value escalation (the mid-level fusion) and the coarse-to-fine framework (the high-level fusion). Two examples are provided for feature selection to demonstrate its ideas, those are, optimal descriptor selection and improved classifier selection.

Feature extraction plays a key role in object recognition because it is the first and also the most important step. If we consider the overall object recognition process, machine learning tools are to serve the purpose of finding distinctive features from the visual data. Given distinctive features, object recognition is readily available (e.g., a simple threshold function can be used to classify feature descriptors). The proposal of Local N-ary Pattern (LNP) texture features contributes to both feature extraction and texture classification. The distinctive LNP feature generalizes the texture feature extraction process and improves texture classification. Concretely, the local binary pattern (LBP) is the special case of LNP with $n = 2$ and the texture spectrum is the special case of LNP with $n = 3$. The proposed LNP representation has been proven to outperform the popular LBP and one of the LBP's most successful extension - local ternary pattern (LTP) for texture classification.

Acknowledgements

I wish to thank Associate Professor Qiang WU and Professor Xiangjian HE, my principle and co-supervisors, for their many suggestions and constant encouragement, help and support during my candidature. I gratefully acknowledge the invaluable discussion with them and I have the honour of studying and working with them in the past four years and nine months which is stamped indelibly in my life.

I appreciate the support of the International Research Scholarship (IRS) provided by the Faculty of Engineering and Information Technology (FEIT), University of Technology, Sydney (UTS). I appreciate the financial support for my living and for attending the international conferences received from the FEIT and the UTS Vice-Chancellor's Conference Fund.

I wish to thank my fellow colleagues and the staff of the faculty for providing various assistance for the completion of this research work. In particular, Wenjing Jia, Min Xu for their invaluable help and support.

Last but not least, I would like to thank my parents for their understanding and support. This thesis could not have been completed without their encouragements and financial assistance.

Contents

CERTIFICATE OF ORIGINAL AUTHORSHIP i

Abstract iii

Acknowledgements v

Contents vi

List of Figures x

List of Tables xiv

Abbreviations xv

1 Introduction 1
- 1.1 Background Information . . . 2
  - 1.1.1 General Object Recognition Framework with Supervised Learning Methods 3
  - 1.1.2 Typical Classification Methods . . . 5
    - 1.1.2.1 Boosting Method and AdaBoost Classifier . . . 5
    - 1.1.2.2 Naive Bayes Classifier . . . 8
    - 1.1.2.3 SVM Classifier . . . 9
  - 1.1.3 Example Datasets in Object Recognition . . . 11
  - 1.1.4 Evaluation Methods and Measurements . . . 12
  - 1.1.5 Summary . . . 15
- 1.2 List of Publications . . . 16
  - 1.2.1 Journal Articles . . . 16
  - 1.2.2 Conference Proceedings . . . 16
- 1.3 Thesis Organization . . . 17

2 Summary of Feature Extraction Methods 18
- 2.1 Over-complete Template Based Feature Description . . . 19
  - 2.1.1 Haar-like Feature for Face Detection . . . 19
  - 2.1.2 Histogram Distance of Haar Regions for Object Recognition . . . 22
  - 2.1.3 Edgelet Features for Pedestrian Detection . . . 23
- 2.2 Sparse Key-point Based Feature Description . . . 26

2.2.1 Image Matching Using Scale Invariant Feature Transform 26
2.3 Dense Spatial Sub-block Based Feature Description 32
2.3.1 Dense SIFT Feature for Image Alignment and Face Recognition 32
2.3.2 Histogram of Oriented Gradient Feature for Pedestrian Detection 36
2.3.3 Density Variance Feature for License Plate Detection 42
2.4 Hybrid Filter Response Based Feature Description 43
2.4.1 Object Bank Representation for Scene Classification 43
2.5 Kernel Based Feature Description 45
2.5.1 Locally Adaptive Regression Kernels for Car Detection 46
2.6 Summary 48

3 Feature Fusion for Object Recognition 51
3.1 Concatenation of Histogram Distances Between Intensity Histograms of Different Spatial Sub-blocks 52
3.1.1 Exploring Relationships Between Different Spatial Sub-blocks 53
3.1.2 Correlated Histogram of Haar Regions 54
3.1.2.1 Original HDHR Distance 55
3.1.2.2 Minkowski City Block L_1 Manhattan Distance 56
3.1.2.3 Minkowski Euclidean L_2 Euclidean Distance 56
3.1.3 Experimental Results 57
3.2 Multi-scale Mid-level Fusion Based on Escalating Confidence Values 60
3.2.1 Multi-scale Gaussian Smoothing 61
3.2.2 Adaptive Combination of Weak Learners Based on Edgelet 63
3.2.3 Concatenation of Confidence Values Across Multi-scale Spaces 66
3.2.4 Experimental Results 67
3.3 Coarse-to-fine Fusion Based on Fast Region of Interest Selection and Discriminative Learning 69
3.3.1 Fast Region of Interest Selection in Coarse Stage 70
3.3.2 Boosting Distinctive DSIFT Features in Fine Stage 71
3.3.3 Experimental Results 74
3.4 Summary 76

4 Feature Selection for Image Classification and Object Recognition 79
4.1 A Compact Descriptor Based on Optimal Max Pooling Orientation 80
4.1.1 A Simplified Strip-like Spatial Configuration 82
4.1.2 Selecting Optimal Strip-like Partition Scheme Based on SVM Classification Results 85
4.1.3 Experimental Results 87
4.2 An Improved Classifier which Requires Less Training Examples 95
4.2.1 An Observation from Using Insufficient Training Examples 96
4.2.2 An Improved AdaBoost Algorithm with Spatial Constraints 97
4.2.3 Experimental Results 103
4.3 Summary 105

5 Local N-ary Pattern for Texture Image Classification 109
5.1 More Distinctive Texture Features in High Dimension Feature Space 110
5.1.1 Mapping Local Patterns to Integer Values 111

5.1.2 Generalization of Feature Extraction to The Bachet De Meziriac Weight Problem . . . 114
5.1.3 Distinctive Local Patterns of The $N-$ry Coding Scheme . . . 120
5.1.3.1 Distinctiveness of LNP Compared to LTP . . . 120
5.1.3.2 An Observation of The Statistical Characteristics of The Local Nary Representation . . . 121
5.1.3.3 Rotation Invariant and Uniform Patterns . . . 125
Uniform Patterns . . . 126
Rotation Invariant Patterns . . . 126
5.1.3.4 Texture Classification . . . 127
One-against-rest SVM Classifier . . . 127
Nearest Neighbor Classifier . . . 128
5.1.4 Experimental Results . . . 128
5.2 Summary . . . 133

6 Conclusions and Future Works 134
6.1 Conclusions . . . 134
6.1.1 Improved Object Recognition Based on Feature Fusion . . . 135
6.1.1.1 Concatenating Distances Between Intensity Histograms of Different Spatial Sub-blocks . . . 135
6.1.1.2 Multi-scale Mid-level Fusion Based on Escalating Confidence Values . . . 136
6.1.1.3 Coarse-to-fine Fusion Based on Fast Region of Interest Selection and Distinctive Features . . . 138
6.1.2 Improved Object Recognition Based on Feature Selection . . . 139
6.1.2.1 A Compact Descriptor Based on Optimal Max Pooling Orientation 140
6.1.2.2 A Classifier which Requires Less Training Examples . . . 141
6.1.3 Improved Texture Classification Based on LNP . . . 142
6.2 Future Works . . . 143

A The Generalized Bachet De Meziriac Weight Problem 146
A.1 A Generalization to The BMWP . . . 146
A.2 Proof of Proposition 1 (Proof by Contradiction) . . . 147
A.2.1 Details of The Contradition . . . 148
A.3 Proof of Proposition 2 (Proof by Mathematical Induction) . . . 148
A.3.1 When $n-1$Is An Even Number . . . 148
A.3.1.1 Basis . . . 148
A.3.1.2 Inductive Step . . . 149
A.3.2 When $n-1$Is An Odd Number . . . 150
A.3.2.1 Basis . . . 150
A.3.2.2 Inductive Step . . . 151

B Some Implementations in Matlab 153
B.1 Recursive SPM Max Pooling . . . 153
B.2 Edgelet Feature Extraction . . . 154

C List of Technique Abbreviation Terms **157**

Bibliography **161**

List of Figures

1.1 A schematic depiction of the AdaBoost VJ cascade 8
1.2 Max-margin hyperplane and margins for an SVM trained with samples from two classes (image courtesy of [1]) 10
1.3 Two example car images for license plate detection 13
1.4 An example of ROC curve (image courtesy of [2]) 14

2.1 A few exemplar Haar-like feature templates [3][4]. Subfigure (A), (B), (C) and (D) are edge templates. Subfigure (E) to (L) are line templates. Subfigure (M), (N), (O) and (P) are center templates. Subfigure (Q), (R), (S) and (T) are diagonal templates 21
2.2 A comparison of detection rates versus false positive regions on the MIT+CMU test set 22
2.3 A few exemplar edgelets [5]. Subfigure (A) is a line edgelet. Subfigure (B), (C) and (D) are $\frac{1}{8}$, $\frac{1}{4}$ and $\frac{1}{2}$ circle edgelets, respectively. Subfigure (E) is a $\frac{1}{4}$ circle symmetric pair 23
2.4 An example of the gradient and the Edgelet response map. Top row from left to right: original image, gradient magnitude map, gradient orientation map. Bottom row from left to right: Edgelet line template 1, Edgelet line template 2, Edgelet map for line template 1, Edgelet map for line template 2 24
2.5 The FB, HS, T and L body parts 25
2.6 An evaluation of the edgelet part based detector (image courtesy of [5]) 26
2.7 Calculate the DoG images (image courtesy of [6]) 27
2.8 Scale space extrema selection (image courtesy of [6]) 27
2.9 An example of SIFT keypoints selection process. Subfigure (a) is the original image, Subfigure (b) is the image with all 832 SIFT keypoints extracted, Subfigure (c) is the image after low contrast keypoints are removed and Subfigure (d) is the image after removal of both low contrast keypoints and poorly located keypoints, such as keypoints located on an edge (images courtesy of [6]) 28
2.10 An illustration of the SIFT feature extraction process, the circle represents the weighing Gaussian window (image courtesy of [6]) 29
2.11 An example illustration of SIFT descriptors for object recognition with occlusion (images courtesy of [6]) 30
2.12 An example illustration of SIFT descriptors for location recognition (images courtesy of [6]) 31
2.13 Emblematic examples of selected DSIFT keypoints in the 3D pyramid in the scale space. White dots are the location of the reference point for DSIFT descriptors and black regions represent the DSIFT patches 33
2.14 An illustration of the*SIFT image*, pixels of a similar structure are grouped together (image courtesy of [7]) 34

2.15 A few example images and their visualized DSIFT images 35
2.16 An example of the various gradient maps. From left to right: original image, horizontal gradient map, vertical gradient map, gradient magnitude map and gradient orientation map (images courtesy of [8]) 37
2.17 An illustration of the HOG feature extraction process. The colored rectangles represent image sub-blocks 37
2.18 The HOG representation of an example test image and the average gradient map (images courtesy of [9]) 39
2.19 Examples images in the MIT CBCL Pedestrian Dataset and INRIA Person Dataset (Images are scaled to fit page width, images in the top two rows are selected from the MIT dataset and images in the bottom two rows are selected from the INRIA dataset) 40
2.20 A comparison of 100, 300 and 500 images averaged on the MIT pedestrian dataset and INRIA human dataset 41
2.21 Sub-block geometry for density variance features (images courtesy of [10]) 42
2.22 A comparison between GIST [11], SIFT-SPM [12] and OB representations (image courtesy of [13]) 43
2.23 Illustration of the OB representation 44
2.24 An example usage of the ‘kernel trick’. Data points not linearly separable in the two dimensional space become linearly separable in a three dimensional space . . . 46
2.25 A graph illustration of the LARK feature extraction process, $p = 3$ in this example 47
2.26 Summary of descriptors relevant to this thesis (descriptors proposed in the author’s publications are highlighted in bold font) 49

3.1 Pairwise comparisons of histograms between different sub-blocks. The intensity histogram extracted from the sub-block filled by horizontal lines is compared to that of the sub-block filled by vertical lines and a distance is calculated for each comparison 53
3.2 ROC curves on MIT CBCL face dataset 58
3.3 ROC curves on MIT CBCL pedestrian dataset 59
3.4 ROC curves on MIT CBCL car dataset 59
3.5 An overview of the mid-level fusion framework 61
3.6 An example of the multi-scale image processing (images courtesy of the INRIA Human Dataset [9]) 63
3.7 Averages images for some objects (generated by Starstax [14]) 64
3.8 Some example images from MIT face (top row on the left), UIUC car (top row on the right) and MIT car (bottom row) 65
3.9 An illustration of stick segments selected under different scales. Each stick segment represents one weak hypothesis in the AdaBoost algorithm, the group of stick segments altogether forms the strong hypothesis 66
3.10 SVM ROC vs. AdaBoost ROC 68
3.11 An overview of the cascaded fusion framework 70
3.12 A schematic depiction of the Softcascade algorithm 73
3.13 Some examples of the license plates used in our experiments [10] 75

4.1 An illustration of the three-level SPM partition scheme and this $1+4+16 = 21$ cells 82
4.2 A brief illustration of the LWOB image representation using strip-like partition . . 83

4.3 Example max pooling of OB using SPM on the response map. The global max response value, 65 (represented by a circle), appears three times in the feature representation, other redundant response values which appear twice are 43 (represented by a rectangle), 57 (represented by a triangle) and 61 (represented by a cross). In this example, the feature representation is {65, 43, 65, 57, 61, 43, 31, 50, 60, 33, 27, 65, 47, 29, 54, 38, 49, 11, 57, 41, 61} 84

4.4 Some misclassified examples in Scene 15 dataset. Top row from left to right: bedroom misclassified as livingroom, industrial misclassified as suburb, industrial misclassified as tallbuilding and industrial misclassified as store (by LWOB). Top-middle row from left to right: kitchen misclassified as livingroom (LWOB) or bedroom (OB), livingroom misclassified as bedroom, coast misclassified as highway and coast misclassified as opencountry. Bottom-middle row from left to right: forest misclassified as mountain (OB), highway misclassified as insidecity, mountain misclassified as opencountry and opencountry misclassified as mountain. Bottom row from left to right: street misclassified as insidecity, tallbuilding misclassified as insidecity (OB), office misclassified as store and store misclassified as industrial (OB) 90

4.5 Some misclassified examples in UIUC sport. Tow row from left to right: croquet misclassified as bocce or polo, badminton misclassified as bocce, bocce misclassified as snowboarding. Bottom row left: RockClimbing misclassified as snowboarding. Bottom row middle and right: rowing misclassified as sailing 91

4.6 Some misclassified examples in Kylberg texture. Top row from left to right: linseeds1 misclassified as sesameseeds1, rug1 misclassified as ceiling1, rug1 misclassified as oatmeal1 and sand1 misclassified as stone1. Bottom row from left to right: stone2 misclassified as stone3, stone2 misclassified as sand1, stone3 misclassified as wall1 and wall1 misclassified as stone1 91

4.7 Comparison on time costs vs. mean classification accuracy in Scene 15 dataset. Best to be viewed in color with magnification. In each figure, blue bars correspond to OB and red bars correspond to the LWOB 92

4.8 Comparison on time costs vs. mean classification accuracy in UIUC Sport dataset. Best to be viewed in color with magnification. In each figure, blue bars correspond to OB, red bars correspond to the LWOB 92

4.9 Comparison on time costs vs. mean classification accuracy in Kylberg Texture dataset. Best to be viewed in color with magnification. In each figure, blue bars correspond to OB, red bars correspond to the LWOB 93

4.10 An illustration of diversity comparison between features. Feature f_3 is more preferable than f_3' because f_3 is more diverse than f_3'. In comparison, the image patch corresponding to feature f_3' significantly overlaps with that of f_2. Consequentially, although the E value for both f_3 and f_3' are the same, the $\frac{A}{G}$ value for f_3 is smaller than that of f_3'. Therefore, f_3 is more preferable than f_3' 99

4.11 Examples of the UIUC Car training subset 103

4.12 Example of the UIUC Car single scale test subset 103

4.13 Example of the UIUC Car multiple scale test subset 104

4.14 Performance comparison between constraint optimization boosting and original boosting on UIUC car single scale test subset (best viewed in color) 105

4.15 Spatial locations of features of five detector models 106

5.1 Texture classes are difficult to discriminate (Images courtesy of [15]). The first and the fourth image (from left to right) in top row belong to the same texture class, while all other images belong to different texture classes despite that they are visually similar. Those seven images in the bottom row belong to three texture classes 111
5.2 LBP feature extraction process . . . 111
5.3 LTP feature extraction process . . . 113
5.4 Illustration of original image and its LBP, LTP and LNP ($n = 3$) representations (best viewed in color) . . . 121
5.5 The LNP (when $n = 3$) is able to discriminate subtle differences between patterns . 122
5.6 Locations of transitions for two exemplary patterns, a three-transition pattern is on the left and a five-transition pattern is on the right . . . 126
5.7 Some example images in the Brodatz texture dataset. First row from left to right: pressedcl, raffia, raffia2, d10, ice, woodgrain, water, ricepaper. Second row from left to right: d51, d95, bark, beachsand, beans, burlap, d4, d5 . . . 129
5.8 Some example images in the CUReT 92 gray image subset. First row from left to right: sample 3, sample 7, sample 13, sample 15, sample 23, sample 30. Second row from left to right: sample 35, sample 38, sample 55, sample 57, sample 58, sample 60 . . . 129
5.9 Some example images in the UIUC texture dataset . . . 130
5.10 Some example images in the Outex texture dataset. First row from left to right: Sample 1 to Sample 10 are in the TC10 subset. Second row from left to right: Sample 11 to Sample 20 are in the TC12 subset . . . 130
5.11 An overview of the TC10 and TC12 subsets in the Outex Database . . . 131

List of Tables

1.1 A list of classification algorithms and their references (Boosting, Naive Bayes and Support Vector Machine classifiers are adopted in this thesis) 4
1.2 Object recognition datasets used in this thesis 12
1.3 The various values for the parameter l and their influences on the sensitivity and the FPR 14

2.1 Definition of all symbols in (2.4) 24

3.1 The relationship between the correlation index i and indices of histograms a and b . 55
3.2 Detailed parameters of those three subsets (face, pedestrian and car) 57
3.3 Speed comparison between COHD and HOG on feature extraction (in seconds) . . 59
3.4 SVM Parameters 67
3.5 Performance comparison of global (coarse) feature level detectors 75
3.6 Performance comparison with [10] and [16] 76
3.7 Detector complexity comparison with previous work [10][16] 76

4.1 Confusion matrix (in percentage) for Scene 15 dataset (numbers for LWOB are in bold font, and numbers for OB are in italic font. Zeros are omitted for brevity)* . . 89
4.2 Confusion matrix (in percentage) for UIUC Sport dataset. Numbers for LWOB are in bold font, and numbers for OB are in italic font. Zeros are omitted for brevity** . 89
4.3 Confusion matrix (in percentage) for Kylberg Texture dataset. Numbers for LWOB are in bold font, and numbers for OB are in italic font. Zeros are omitted for brevity*** 94

5.1 LBP representations follow binomial distribution 123
5.2 Details of those 9 groups of LBP representations in binary form 123
5.3 LBP representations are permutations of binary digits 123
5.4 LNP representations follow a trinomial distribution when $n = 3$ 124
5.5 Summary of distributions for LNP representations 125
5.6 A brief summary of parameters of three non-RIU2 datasets 129
5.7 A brief summary of parameters of the TC10 and TC12 RIU2 datasets 130
5.8 A comparison of LNP and its variant descriptors' dimensionality under different n values 131
5.9 Average of mean classification accuracies (in percentage) for non-RIU2 datasets . . 132
5.10 Mean classification accuracies (in percentage) for RIU2 datasets 132

Abbreviations

a.k.a	**a**lso **k**nown **a**s
e.g.	**e**xemplī **g**rātiā (for example)
i.e.	**i**d **e**st (that is)
max	**max**imum
min	**min**imum
Q.E.D.	**Q**uod **E**rat **Demonstrandum** (which had to be proven)
s.t.	**s**ubject **t**o
vs.	**v**er**s**us

To my paternal grandfather and maternal grandfather.