NEW DESCRIPTOR FOR OBJECT DETECTION USING AN IMPROVED
ENSEMBLE-BASED TECHNIQUE

AMIN MOHAMED AHSAN

UNIVERSITI TEKNOLOGI MALAYSIA
NEW DESCRIPTOR FOR OBJECT DETECTION USING AN IMPROVED ENSEMBLE-BASED TECHNIQUE

AMIN MOHAMED AHSAN

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

MAY 2015
To my beloved parents, my wife, children, brothers and sisters
ACKNOWLEDGMENTS

All praise and thanks are due to Allah who "Taught man that which he knew not," and peace and blessings of Allah be upon our prophet, Mohammad and upon all his family and companions. All thanks to Allah for graces and for giving me the strength and endurance to complete this research

During PhD journey, I was in contact with many people, scholars and academicians colleagues. They have contributed in different ways towards my understanding and thoughts. First, I would like to express my high appreciation to my supervisor, Prof. Dr. Dzulkifli Bin Mohamad, for encouragement, guidance, comments and his support which participated actively in the completion of this work.

My sincere appreciation also extends to Dr. Faisal Alsamet for his unlimited help, Dr. Murad Qasem and Dr. Redhwan Shaddad for their support.
ABSTRACT

Object detection is an essential process for further tasks including, but not limited to, object and event detection, object tracking, object recognition, video indexing, motion estimation, image restoration, image registration, image retrieval, and reconstruction of 3D scene. In the recent past, interest point detectors and their descriptors, as local features, have received a great interest in computer vision areas and technologies. These types of features have shown their robustness against different types of deformation due to geometric transformation, photometric transformation and other disturbances. Therefore, they are more accurate and stable than the global ones. Among all interest point detectors and descriptors, the Scale Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) are considered as the most common methods that receive interest from researchers in terms of usage and development; but, getting more accurate, invariant and fast descriptor is still needed. Matching technique is often used to recognize the object based on such features; however, it is not proper for some applications such as searching for an isolated object and it is difficult to be used in object category recognition or to recognize the part-based object. Therefore, learning-based technique, that has been proven to be an effective method in object detection, can be used to overcome the previously mentioned challenges. However, the object required to be detected usually represents a small ratio compared to non-object that causes an imbalanced data problem. The aim of this study is to design and develop an effective model for object detection that is faster, more accurate and it can manage aforementioned challenges. To achieve this goal first, new fast and an accurate descriptor is introduced based on interest points; second, an effective classification method, that mitigates the effect of imbalanced data, is designed based on developed ensemble classifiers; third, an updating scheme of interest point detector is presented to speed up the object detection system. Results show that the proposed features are faster and more invariant than the most common interest-point-based features. The developed technique based on ensemble classifiers produces notable results in terms of accuracy and False Positive rate compared to the traditional one. The speed of object detection system has increased by 30% in average based on the proposed scheme.
ABSTRAK

Pengesanan objek adalah suatu proses yang penting dalam tugaslan lanjut termasuk pengesanan objek dan acara, penjeakan objek, pengecaman objek, mengindek video, anggaran pergerakan, restorasi imej, pendaftaran imej, penemuan kembali imej and rekonstruksi adegan 3D. Sebelum ini, pengesan titik minat dan pemarialnya, iaitu ciri-ciri sedia ada telah menarik minat yang mendalam di dalam bidang visi komputer dan teknologi. Ciri-ciri ini telah mempermanen kecekenal terhadap beberapa jenis kecacatan yang disebabkan oleh transomasi geometri, transomasi fotometri dan gangguan yang lain. Oleh itu, ciri-ciri ini lebih tepat dan stabil daripada ciri-ciri global yang lain. Antara semua pengesan titik minat dan pemarialnya, Ciri Transform Skala Invarian (SIFT) dan Ciri-Ciri Kelajuan Teguh (SURF) dianggap sebagai dua kaedah sepunya yang paling mendapat perhatian pengkaji dari aspek penggunaan dan pembangunan tetapi penemuan ciri-ciri kategori objek yang lebih cekal dan pantas masih diperlukan. Teknik pemadanan sering digunakan untuk mengenal pasti objek berdasarkan ciri-ciri tersebut. Namun, ia tidak sesuai untuk beberapa aplikasi seperti gelintar objek yang terpencil dan sukar untuk digunakan semasa pengecaman kategori objek atau pengecaman objek yang berdasarkan bahagian-bahagian. Justeru, teknik berdasarkan pembelajaran yang telah dibuktikan sebagai suatu kaedah yang berkesan untuk mengecam objek boleh digunakan untuk mengatasi cabaran-cabaran yang disebut sebelum ini. Walau bagaimanapun, objek yang perlu dikesan lazimnya melambangkan suatu nisbah yang kecil berbanding dengan benda-benda bukan berbentuk objek yang menimbulkan masalah ketidakseimbangan data. Tujuan kajian ini adalah untuk mereka bentuk dan membangunkan satu model yang berkesan untuk mengesan objek dengan lebih pantas, lebih tepat dan dapat mengurus cabaran-cabaran yang disebut tadi. Untuk mencapai matlamat ini, pertama, pengesan yang pantas dan tepat diperkenalkan berdasarkan titik-titik berkepentinan; kedua, satu kaedah klasifikasi yang berkesan dan mampu mengurangkan kesan ketidakseimbangan data yang direka bentuk berdasarkan pengklasifikasi ensembel yang dibangunkan; ketiga, dikemukakan suatu skema yang dapat mengemaskini pengesan titik minat untuk mempercepatkan sistem pengecaman objek. Dapatkan kajian menunjukkan bahawa ciri-ciri yang dicadangkan adalah lebih pantas dan lebih tepat berbanding ciri-ciri yang berdasarkan titik kepentinan yang sepunya. Teknik yang dibangunkan berdasarkan pengklasifikasi ensembel telah menghasilkan keputusan yang penting berkaitan kejadian dan kadar positif palsu berbanding dengan teknik tradisional. Namun, kelajuan sistem pengesanan objek telah meningkat 30% secara purata berdasarkan skema yang dicadangkan.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iv</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xvi</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xxiii</td>
<td></td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xxv</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Problem Background</td>
<td>2</td>
</tr>
<tr>
<td>1.3</td>
<td>Problem Statement</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>Research Questions</td>
<td>7</td>
</tr>
<tr>
<td>1.5</td>
<td>Research Aim</td>
<td>8</td>
</tr>
<tr>
<td>1.6</td>
<td>Research Objectives</td>
<td>8</td>
</tr>
<tr>
<td>1.7</td>
<td>Research Scope and Assumptions</td>
<td>9</td>
</tr>
<tr>
<td>1.8</td>
<td>Research Significance</td>
<td>10</td>
</tr>
<tr>
<td>1.9</td>
<td>Summary of Research Contributions</td>
<td>11</td>
</tr>
<tr>
<td>1.10</td>
<td>Terms Definitions</td>
<td>11</td>
</tr>
<tr>
<td>1.11</td>
<td>Thesis Outline</td>
<td>13</td>
</tr>
</tbody>
</table>
2 LITERATURE REVIEW

2.1 Introduction
2.2 Object Detection
  2.2.1 Features
    2.2.1.1 Interest Points Detectors
    2.2.1.2 Interest Points’ Descriptors
      2.2.1.2.1 Floating-point-based Descriptors
      2.2.1.2.2 Binary-based Descriptors
    2.2.1.3 Interest Point Detectors used and their Descriptors
      2.2.1.3.1 SIFT Points’ Detector and Descriptor
      2.2.1.3.2 SURF Points’ Detector and Descriptor
  2.2.2 Object Recognition
    2.2.2.1 Supervised Learning Algorithms
    2.2.2.2 Technique Used
      2.2.2.2.1 Back-propagation Neural Network
    2.2.2.2.2 Related Issues
  2.2.3 Object Localization
  2.3 Ensemble-based Recognition
  2.4 Data sets
  2.5 Performance Evaluation
  2.6 Related Works
    2.6.1 Object Detection Frameworks
    2.6.2 Features
    2.6.3 Recognition
  2.7 Discussion

3 RESEARCH METHODOLOGY

3.1 Introduction
3.2 Study Framework 75
3.3 Object Detection Datasets 78
  3.3.1 Faces Dataset 78
    3.3.1.1 Training Data 78
    3.3.1.2 Testing Data 81
      3.3.1.2.1 Labeled Data 81
      3.3.1.2.2 Unlabeled Data 83
  3.3.2 Cars Dataset 84
    3.3.2.1 Training Data 84
    3.3.2.2 Testing Data 86
      3.3.2.2.1 Labeled Data 86
      3.3.2.2.2 Unlabeled Data 88
3.4 Proposed Methods 89
3.5 Experimental Design 91
3.6 Performance Evaluation 92
3.7 Summary 94

4 NESTED-WINDOWS OF HISTOGRAM OF ORIENTED GRADIENTS DESCRIPTOR 95
4.1 Introduction 95
4.2 Methods 96
  4.2.1 Motivation 96
  4.2.2 SIFT and SURF Descriptors 97
    4.2.2.1 SIFT Descriptor 97
    4.2.2.2 SURF Descriptor 98
  4.2.3 The Proposed Descriptor 99
  4.2.4 Single-based Learning Technique 106
    4.2.4.1 Learning Phase 106
    4.2.4.2 Testing Phase 109
4.3 Experimental Design 110
4.4 Experimental Results 111
  4.4.1 Classification Performance 113
    4.4.1.1 Faces Classification 114
4.4.1.2   Cars Classification 116
4.4.2    Localization Performance 118
4.4.2.1   Faces Localization 119
4.4.2.2   Cars Localization 120
4.4.3    Invariance 121
4.4.3.1   Invariance to Rotation 121
4.4.3.2   Invariance to Scale 124
4.5    Time Consumption 128
4.6    Discussion 130
4.6    Summary 135

5    AN IMPROVED ENSEMBLE-BASED TECHNIQUE
FOR OBJECT CLASSIFICATION 136
5.1    Introduction 136
5.2    Methods 137
5.2.1    Motivation 137
5.2.2    An Ensemble-based Developed Technique 138
5.2.2.1    Learning Phase 140
5.2.2.2    Testing Phase 142
5.3    Experimental Design 144
5.4    Experimental Results 144
5.4.1    Classification Performance 145
5.4.1.1    Faces Classification Performance 145
5.4.1.2    Cars Classification Performance 150
5.4.2    Localization Performance 155
5.4.2.1    Faces Localization Performance 155
5.4.2.2    Cars Localization Performance 156
5.4.3    Time Consumption 157
5.5    Discussion 163
5.6    Summary 170
## AN IMPROVED SCHEME OF AN INTEREST POINT DETECTOR

6.1 Introduction 171
6.2 Motivation 172
   6.2.1 Methods 172
   6.2.2 A Developed Scheme for Speeding up The Object Detection System 172
6.3 Experimental Results 179
6.4 Discussion 185
6.5 Summary 188

## CONCLUSIONS

7.1 Introduction 189
7.2 Summary of Research Achievements 190
7.3 Research Contributions 191
7.4 Future Works 194
7.5 Research Challenges 195

## REFERENCES

Appendices A-C 210-249
### LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison between the global and local features</td>
<td>21</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparison among some existing points’ detector regarding some important criterion</td>
<td>25</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison among the most common algorithms (Kotsiantis, et al., 2007)</td>
<td>40</td>
</tr>
<tr>
<td>2.4</td>
<td>Summary of some existing studies in regard to localization performance and time consuming</td>
<td>48</td>
</tr>
<tr>
<td>2.5</td>
<td>Notes on existing solutions of imbalanced data</td>
<td>53</td>
</tr>
<tr>
<td>2.6</td>
<td>INRIA person dataset</td>
<td>56</td>
</tr>
<tr>
<td>2.7</td>
<td>Summary of some common datasets used in object detection field</td>
<td>57</td>
</tr>
<tr>
<td>2.8</td>
<td>The measurements used in some existing studies</td>
<td>61</td>
</tr>
<tr>
<td>2.9</td>
<td>Some proposed frameworks in literature</td>
<td>63</td>
</tr>
<tr>
<td>3.1</td>
<td>The relations among phases, study questions, study objectives, methods, chapters, and the evaluation</td>
<td>90</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison among SIFT, SURF, and NHOG regarding some criterion</td>
<td>105</td>
</tr>
<tr>
<td>4.2</td>
<td>Summary of Datasets used</td>
<td>112</td>
</tr>
<tr>
<td>4.3</td>
<td>Number of nodes of hidden layer for each descriptor</td>
<td>113</td>
</tr>
<tr>
<td>4.4</td>
<td>Comparison between SURF and NHOG in terms of classification performance on Faces dataset</td>
<td>114</td>
</tr>
<tr>
<td>4.5</td>
<td>Comparison between SIFT and NHOG in terms of classification performance on Faces dataset</td>
<td>114</td>
</tr>
<tr>
<td>4.6</td>
<td>Comparison between SURF and NHOG in terms of classification performance on Cars dataset</td>
<td>116</td>
</tr>
<tr>
<td>4.7</td>
<td>Comparison between SIFT and NHOG in terms of</td>
<td></td>
</tr>
</tbody>
</table>
classification performance on Cars dataset

4.8 Comparison between SURF and NHOG in terms of point localization on faces dataset evaluated by Precision measurement

4.9 Comparison between SIFT and NHOG in terms of point localization on faces dataset evaluated by Precision measurement

4.10 Comparison between SURF and NHOG in terms of point localization on cars dataset evaluated by Precision measurement

4.11 Comparison between SIFT and NHOG in terms of point localization on cars dataset evaluated by Precision measurement

4.12 Comparison between SURF and NHOG regarding to rotation changes

4.13 Comparison between SIFT and NHOG regarding to rotation changes

4.14 Comparison between SURF and NHOG regarding to Scale changes

4.15 Comparison between SIFT and NHOG regarding to Scale changes

4.16 Time consumption comparison among SURF, SIFT, and NHOG descriptor based on single classifier

4.17 Paired Samples Test on faces dataset, single-based classifier

4.18 Paired Samples Statistics on faces dataset, single-based classifier

4.19 Paired Samples Test on cars dataset, single-based classifier

4.20 Paired Samples Statistics on cars dataset, single-based classifier

5.1 Comparison between SURF and NHOG in terms of classification performance on Faces dataset for E1 and E2 environments

5.2 Comparison between SIFT and NHOG in terms of classification performance on Faces dataset for E1 and E2 environments

5.3 Comparison between SURF and NHOG in terms of classification performance on faces dataset based on
5.4 Comparison between SIFT and NHOG in terms of classification performance on faces dataset based on ensemble learning technique

5.5 Comparison between SURF and NHOG in terms of classification performance on cars dataset for E1 and E2 environments

5.6 Comparison between SIFT and NHOG in terms of classification performance on cars dataset for E1 and E2 environments

5.7 Comparison between SURF and NHOG in terms of classification performance on cars dataset based on ensemble learning technique

5.8 Comparison between SIFT and NHOG in terms of classification performance on cars dataset based on ensemble learning technique

5.9 Comparison between SURF and NHOG in terms of point localization on faces dataset evaluated by Precision measurement based on ensemble classifiers

5.10 Comparison between SIFT and NHOG in terms of point localization on faces dataset evaluated by Precision measurement based on ensemble classifiers

5.11 Comparison between SURF and NHOG in terms of point localization on cars dataset evaluated by Precision measurement based on ensemble classifiers

5.12 Comparison between SIFT and NHOG in terms of point localization on cars dataset evaluated by Precision measurement based on ensemble classifiers

5.13 Time consuming, SURF vs. NHOG and SIFT vs. NHOG applied on faces dataset

5.14 Time consuming, SURF vs. NHOG and SIFT vs. NHOG applied on cars dataset

5.15 Paired Samples Test on faces dataset, Single-based vs. Ensemble-based techniques

5.16 Paired Samples Statistics on faces dataset for Single-based and Ensemble-based techniques
5.17 Paired Samples Test on cars dataset, Single-based vs. Ensemble-based techniques 169
5.18 Paired Samples Statistics on cars dataset for Single-based and Ensemble-based techniques 169
6.1 TPP, FP, TP, and Time consuming regarding the octave and hidden layer settings 176
6.2 TPP, FP, TP, and Time consuming using NHOG+ with different hidden layer settings 177
6.3 The average of time (in seconds) and the improvement percentage for NHOG+ over NHOG applied on faces and cars datasets 179
6.4 TPP, FP, TP, and Precision of NHOG+ applied on faces and cars datasets 182
6.5 Paired Samples Test regarding time consuming, NHOG vs. NHOG+ 185
6.6 Paired Samples Statistics regarding time consuming, NHOG +vs. NHOG 186
6.7 Paired Samples Test regarding Precision metric, NHOG vs. NHOG+ 187
6.8 Paired Samples Statistics regarding Precision metric, NHOG +vs. NHOG 187
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Scenario leading to the study problem</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>The most common object detection challenges</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>Taxonomy of the main steps and elements of object detection system</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Taxonomy of features in object detection</td>
<td>19</td>
</tr>
<tr>
<td>2.4</td>
<td>Most common points’ detectors methods</td>
<td>22</td>
</tr>
<tr>
<td>2.5</td>
<td>Convolved Gaussian (left), DOG (right) proposed by Lowe (2004)</td>
<td>30</td>
</tr>
<tr>
<td>2.6</td>
<td>Maxima and minima of DOG, Lowe (2004)</td>
<td>31</td>
</tr>
<tr>
<td>2.7</td>
<td>Interest point descriptor, Lowe (2004)</td>
<td>32</td>
</tr>
<tr>
<td>2.8</td>
<td>Computing the upright rectangle area using integral Image Bay et al. (2008)</td>
<td>34</td>
</tr>
<tr>
<td>2.9</td>
<td>LOG Approximation (top row), Bay formula (2.13) (bottom row)</td>
<td>35</td>
</tr>
<tr>
<td>2.10</td>
<td>Scale-Space: SIFT (left), SURF (right)</td>
<td>35</td>
</tr>
<tr>
<td>2.11</td>
<td>Assigning the Orientation, Bay et al. (2008)</td>
<td>36</td>
</tr>
<tr>
<td>2.12</td>
<td>Components of SURF Descriptor</td>
<td>37</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.13</td>
<td>Taxonomy of machine learning techniques</td>
<td>39</td>
</tr>
<tr>
<td>2.14</td>
<td>Standard architecture of Back-propagation Neural Network</td>
<td>43</td>
</tr>
<tr>
<td>2.15</td>
<td>Steps of BP neural network</td>
<td>45</td>
</tr>
<tr>
<td>2.16</td>
<td>Taxonomy of imbalanced data solutions</td>
<td>51</td>
</tr>
<tr>
<td>2.17</td>
<td>Bi-level filters for Center-Surround (Agrawal, et al., 2008)</td>
<td>65</td>
</tr>
<tr>
<td>2.18</td>
<td>Proposed approach for face detection (D. Kim and Dahyot, 2008)</td>
<td>68</td>
</tr>
<tr>
<td>2.19</td>
<td>Fusion schemes proposed by (Ludwig, et al., 2009). (a) scheme1, (b) scheme2, (c) scheme3</td>
<td>70</td>
</tr>
<tr>
<td>2.20</td>
<td>Cascade classifiers structure (Wang, et al., 2012)</td>
<td>71</td>
</tr>
<tr>
<td>3.1</td>
<td>The Study Framework</td>
<td>76</td>
</tr>
<tr>
<td>3.2</td>
<td>Some examples of Caltech 101 “faces” dataset, the original images</td>
<td>79</td>
</tr>
<tr>
<td>3.3</td>
<td>Some examples of an isolated faces clipped from Caltech 101 “faces” dataset, for training</td>
<td>79</td>
</tr>
<tr>
<td>3.4</td>
<td>Indoor examples from SUN397 dataset, for training</td>
<td>80</td>
</tr>
<tr>
<td>3.5</td>
<td>Outdoor examples from SUN397 dataset, for training</td>
<td>80</td>
</tr>
<tr>
<td>3.6</td>
<td>Some examples of an isolated faces clipped from Caltech 101 “faces” dataset, for classification test</td>
<td>82</td>
</tr>
<tr>
<td>3.7</td>
<td>Some examples of non-object from SUN397, for classification test</td>
<td>82</td>
</tr>
<tr>
<td>3.8</td>
<td>Some unlabelled images from Caltech 101 “faces”, for localization test</td>
<td>83</td>
</tr>
<tr>
<td>3.9</td>
<td>Some examples of car object used for training</td>
<td>85</td>
</tr>
</tbody>
</table>
3.10 Some examples of non-object (street) used for cars training 85
3.11 Some examples of non-object (open space) used for cars training 86
3.12 Some examples of car object used for classification test 87
3.13 Some examples of non-object used for classification test 87
3.14 Some unlabelled images from VOC2009, for localization test 88
4.1 SIFT descriptor, Lowe (2004) 97
4.2 Orientation assigning, Bay et al. (2008) 98
4.3 SURF Descriptor Components 99
4.4 Points’ detection. (a) original image, (b) detected points 100
4.5 Building up the window around interest point 100
4.6 NHOG features extraction elements: (a) window gradients, (b) nested-windows of gradients, (c) the considered gradients, (d) pyramid representation 103
4.7 The involved process in NHOG 104
4.8 General framework for object detection based on interest points 106
4.9 Training Procedure 108
4.10 Testing Procedure 110
4.11 ROC curve for SURF and NHOG according to hidden layer settings on Faces dataset. (left) H1, (right) H2 115
4.12 ROC curve for SIFT and NHOG according to hidden layer settings on Faces dataset. (left) H1, (right) H2 115
4.13 ROC curve for SURF and NHOG according to hidden layer 117
4.14 ROC curve for SIFT and NHOG according to hidden layer settings on Cars dataset. (left) H1, (right) H2

4.15 Comparison between SURF and NHOG in regards to average of FP along all images at θ

4.16 Comparison between SIFT and NHOG in regards to average of FP along all images at θ

4.17 Comparison between SURF and NHOG in regards to average of FP along all images at different scale value

4.18 Comparison between SIFT and NHOG in regards to average of FP along all images at different scale value

4.19 Time consumption for each image using SURF and NHOG descriptor based on single classifier

4.20 Time consumption for each image using SURF and NHOG descriptor based on single classifier

5.1 Learning and testing phase of the developed technique. (a) learning, (b) testing

5.2 ROC curve for SURF and NHOG according to E1 and E2 environments on Faces dataset with H1 and H2 setting. (top) H1, (bottom) H2

5.3 ROC curve for SIFT and NHOG according to E1 and E2 environments on Faces dataset with H1 and H2 setting. (top) H1, (bottom) H2

5.4 ROC curve for SURF and NHOG according to H1 and H2 applied on Faces dataset based on ensemble classifiers. (left) H1, (right) H2

5.5 ROC curve for SIFT and NHOG according to H1 and H2 applied on Faces dataset based on ensemble classifiers. (left)
ROC curve for SURF and NHOG according to E1 and E2 environments on cars dataset with H1 and H2 setting. (top) H1, (bottom) H2

ROC curve for SIFT and NHOG according to E1 and E2 environments on cars dataset with H1 and H2 setting. (top) H1, (bottom) H2

ROC curve for SURF and NHOG according to H1 and H2 applied on cars dataset based on ensemble classifiers. (left) H1, (right) H2

ROC curve for SIFT and NHOG according to H1 and H2 applied on cars dataset based on ensemble classifiers. (left) H1, (right) H2

Comparison between SURF and NHOG in terms of time consuming for each image using ensemble-based technique applied on faces dataset. (a) H1, (b) H2

Comparison between SIFT and NHOG in terms of time consuming for each image using ensemble-based technique applied on faces dataset. (a) H1, (b) H2

Comparison between SURF and NHOG in terms of time consuming for each image using ensemble-based technique applied on cars dataset. (a) H1, (b) H2

Comparison between SIFT and NHOG in terms of time consuming for each image using ensemble-based technique applied on cars dataset. (a) H1, (b) H2

Comparison between single-based and ensemble-based techniques for SURF and NHOG descriptors on faces dataset in regard to H1 and H2 setting. (top) H1, (bottom) H2
5.15 Comparison between single-based and ensemble-based techniques for SIFT and NHOG descriptors on faces dataset in regards to H1 and H2 setting. (top) H1, (bottom) H2 164

5.16 Comparison between single-based and ensemble-based techniques for SURF and NHOG descriptors on cars dataset in regards to H1 and H2 setting. (top) H1, (bottom) H2 165

5.17 Comparison between single-based and ensemble-based techniques for SIFT and NHOG descriptors on cars dataset in regards to H1 and H2 setting. (top) H1, (bottom) H2 166

6.1 Filters sizes for each octave used in SURF 174

6.2 SURF point detector, the relation between scale and the detected interest points 174

6.3 Initial proposed schemes to speed up point’s detector, a) 4 octaves, b) 3 octaves 175

6.4 Proposed scheme to speed up point’s detector, NHOG+ 177

6.5 Some examples of object localization regarding the octave settings and hidden layer settings, left half H1, right half H2. (a) 4 octaves, (b) 3 octaves, (c) odd octaves. 178

6.6 Comparison between NHOG and NHOG+ in terms of time consumption regarding hidden layer setting applied on faces dataset. (a) H1, (b) H2 180

6.7 Comparison between NHOG and NHOG+ in terms of time consumption regarding hidden layer setting applied on cars dataset. (a) H1, (b) H2 181

6.8 Comparison between NHOG and NHOG+ in terms of FP error regarding hidden layer setting applied on faces dataset. (a) H1, (b) H2 183
6.9 Comparison between NHOG and NHOG+ in terms of FP error regarding hidden layer setting applied on cars dataset.

(a) H1, (b) H2
### LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>B-SIFT</td>
<td>Binary SIFT</td>
</tr>
<tr>
<td>BRIEF</td>
<td>Binary Robust Independent Elementary Features</td>
</tr>
<tr>
<td>BRISK</td>
<td>Binary Robust Invariant Scalable Key-points</td>
</tr>
<tr>
<td>CCR</td>
<td>Correct Classification Rate</td>
</tr>
<tr>
<td>CenSurE</td>
<td>Center Surround Extermas</td>
</tr>
<tr>
<td>CV</td>
<td>Computer Vision</td>
</tr>
<tr>
<td>DOC</td>
<td>Difference of Gaussian</td>
</tr>
<tr>
<td>DR</td>
<td>Detection Rate</td>
</tr>
<tr>
<td>FAST</td>
<td>Features from Accelerated Segment Test</td>
</tr>
<tr>
<td>FLDA</td>
<td>Fishier Linear Discriminated Analysis</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FNR</td>
<td>False Negative Rate</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>FPR</td>
<td>False Positive Rate</td>
</tr>
<tr>
<td>GLOH</td>
<td>gradient location-orientation histogram</td>
</tr>
<tr>
<td>HOG</td>
<td>Histogram of Oriented Gradient</td>
</tr>
<tr>
<td>IM-SURF</td>
<td>Improved SURF</td>
</tr>
<tr>
<td>LOG</td>
<td>Laplacian of Gaussian</td>
</tr>
<tr>
<td>NHOG</td>
<td>Nested-windows of Histogram of Oriented Gradients</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NHOG+</td>
<td>Nested-windows of Histogram of Oriented Gradients plus</td>
</tr>
<tr>
<td>ORD</td>
<td>Oriented FAST and Rotated BRIEF</td>
</tr>
<tr>
<td>PCA-SIFT</td>
<td>Principal Component Analysis SIFT</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale Invariant Feature Transform</td>
</tr>
<tr>
<td>SURF</td>
<td>Speeded Up Robust Feature</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TCB</td>
<td>Totally Corrective Boosting</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TNR</td>
<td>True Negative Rate</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>TPR</td>
<td>True Positive Rate</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>False Positive and True Positive over all images applied on faces and cars datasets using single-based classifier</td>
<td>210</td>
</tr>
<tr>
<td>B</td>
<td>False Positive and True Positive over all images applied on faces and cars datasets using ensemble-based classifier</td>
<td>226</td>
</tr>
<tr>
<td>C</td>
<td>False Positive and True Positive over all images applied on faces and cars datasets using ensemble-based classifier and NHOG+ scheme</td>
<td>242</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Introduction

Computer vision (CV) is an area that includes approaches and techniques to acquire, process, analyze, and understand an image. Moreover, it produces a new form of symbols information as high-dimensional data from the natural world for further decisions (Morris, 2004). The idea of developing this field is to inspire the abilities of human vision to understand an image by machine (Sonka, 2007). Therefore, an image can be understood by seeing it in another form (i.e. numerical or symbolic information), this information can be extracted by using specific models such as geometry, statistic, learning technique or by combining two or more methods (Forsyth and Ponce, 2002). Additionally, CV is a relatively new area and has great relationship with many other fields such as Machine Learning, Physics, Control Systems, and Medical Imaging. Therefore, it is reasonable that CV has been rapidly developed into many kinds of sub-fields (Calonder, 2010a). Sub-fields of computer vision include, but not limited to, reconstruction of scene, object and events detection, object tracking, recognition of an object, video indexing, motion estimation, and image restoration.

One of the most important processes in CV is the object detection because it is essential process for many other applications such as object tracking, object recognition, object categorization, event detection, and searching for object in
sequence of images or video. In addition, detecting and finding a specific object in sequence images or video has become an increasingly important problem with numerous applications, especially with the popularity of smart phones and other devices. That includes searching for people, buildings, landmarks and other interest objects for further analyzing and decisions (Aly, 2011).

1.2 Problem Background

Visual object detection is an extremely troublesome computational issue. The main problem is that each object in the world can cast an unbound number of diverse 2-D images onto the retina as the object's position, pose, lighting, and background changes in respect to the viewer (DiCarlo and Cox, 2007). The variations occurs due to the deformations of non-rigid, and intra-class variability because of shape and other visual properties, also increase the difficulty of tackling this problem (Felzenszwalb et al., 2010). Furthermore, the quality of object detection (detection rate, precision, error) and time consumption add more difficulties to object detection system in which they are trade-off problems.

Full object detection system normally consists of three main elements; features, recognition, and localization. Features can be divided into two types, according to the area (spatial) that they are taken from, and the methods of feature’s acquisition. Features based on spatiality can be global or local (Aly, 2011; Li B. et al., 2011; Pang et al., 2012); while the ways of getting features can be categorized into three types; extraction, selection, composition features (Lillywhite et al., 2012b).

Recognizing an object can be done based on either matching or learning technique. Matching technique recognize the object based on the distance between the features of the target and the template stored. On the other hand, the learning
technique is a way to build a system that mimic the human ability of learning from examples and it can recognize unknown instance based on prior knowledge and experience. Machine learning techniques can be classified into two types; single-based and ensemble-based techniques (Xu et al., 2011; Zhang X. and Cheng, 2012). In single-based method, only one algorithm is used to learn and classify the instance (Xu et al., 2011). On the other hand, the ensemble-based approach uses a several algorithms in a certain arrangement to get the final decision of instance recognition (Ceamanos et al., 2010; Galar et al., 2012; Liu and Wang, 2010; Rahman and Verma, 2011; Verikas et al., 2010; Xu et al., 2011).

Features play an important role in object detection and affect directly on the quality of detection (Yilmaz et al., 2006). For any object detection within an image based on features, object needs to be represented by features as a first process by which can be used later in the classification phase (Aly, 2011). Features can be divided into two types; global and local.

**Global features:** where an image can be represented by extracting features as one dimension which describe the whole image. Information extracted might be histogram of color (Forsyth and Ponce, 2002), or histogram of orientation as in (Dalal and Triggs, 2005). global features are fast and easy in computation which is considered advantage over local features, but they perform worse than local features in term of accuracy (Douze et al., 2009).

**Local features:** where the features that represent image are extracted from a set of regions by specific descriptor. Generally, there are two steps to extract the local features which are feature detection and feature description (Forsyth and Ponce, 2002; Mikolajczyk and Schmid, 2004a). Feature detection involves detecting interest locations in the image (i.e. corners and edges). Feature description is to describe the image patch that surrounds the interest point; this can be done by histograms of gradients or orientation. There are different types of feature detectors; however among the most well-known ones are Difference of Gaussians (DOG) which was
presented by Lowe (2004), Hessian-Affine and Harris-Affine which were introduced by Mikolajczyk and Schmid (2004), and SURF which brought by Bay et al. (2008) and it is relatively new. Similarly, there are many feature descriptors, SIFT and SURF are the most widely used because of their performance. The local features have superior performance over the global ones; but with large memory usage (Douze et al., 2009).

Among all local features, SIFT and SURF are the dominant features over all others in this field in terms of robustness and invariance (Li H. et al., 2013; Pang et al., 2012). However, the high dimensionality and computation complexity of such features consume a large amount of space (memory) and take longer time in recognition, especially against large-scale dataset e.g. SIFT. Besides, SURF descriptor might not make use the full information about the interest point (Li H. et al., 2013); further, SURF has limitation regarding scale changes in case of zoom-out variant (Bay et al., 2008). So, these factors are challenges in some applications especially in real-time and mobile-based applications (Heinly et al., 2012; Li H. et al., 2013; Rister et al., 2013). These challenges give a motivation to introduce new features by converting the original features of SIFT or SURF into a new binary form. Examples of such binary-based features include, Binary Robust Independent Elementary Features (BRIEF) (Calonder et al., 2010b), Oriented FAST and Rotated BRIEF (ORD) (Rublee et al., 2011), and Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011). Binary-based methods introduced features that are faster and easy to be computed, however, they still inaccurate, instable, and less descriptive than the original ones (Heinly et al., 2012; Ni, 2012).

Regarding the object recognition; matching technique is used to find the corresponding points between two images depending on the distances between the accompanying features. It is common technique used in object recognition and image registration (Brown et al., 2011). But, computing the descriptor and using the matching against large database adds more complexity to the recognition system (Gauglitz et al., 2011). It is also difficult to be used in object category recognition or
to recognize the part-based object. Moreover, it is not proper for some applications such as searching for an isolated object (Gauglitz et al., 2011).

As alternative to the matching technique, the learning-based technique, it has proven to be an effective method in object detection (Qiu et al., 2012). It is used successfully in category classification especially discriminative-based learning (Gu and Ren, 2010b), also researchers are paid more attention to design method that learn the interest point features (Li B. et al., 2011). Examples of such techniques are, Artificial Neural Network (ANN), boosting, decision trees, and Support Vector Machines (SVM) (Porikli and Yilmaz, 2012).

Since the data in object detection is an imbalance, this is a big challenge facing the learning-based methods as a single learning because the single-based method expect that all classes have the same distribution (He and Garcia, 2009; Sun et al., 2009), which leads to poorly accuracy of classification (Ceamanos et al., 2010; Galar et al., 2012; He and Garcia, 2009; Sun et al., 2009). Therefore, most conventional (standard or single-based) methods fail to be an effective to fix such problem (Ceamanos et al., 2010; Galar et al., 2012; He and Garcia, 2009).

To tackle this challenge, ensemble-based techniques are proposed to improve the accuracy of the classification (Ceamanos et al., 2010; Galar et al., 2012), and to reduce the effects of imbalanced data on the classification quality (Galar et al., 2012; Zhang X. and Cheng, 2012; Zhang Y. et al., 2012). however, this type of learning has structure complexity and need longer time than single-based learner (Xu et al., 2011).

Based on the afore-stated issues; Figure 1.1 summarizes the advantages, limitation of each parts of object detection system and the scenario that led to the study problem.
Adv.=Advantages, Disadv.=Disadvantages

**Figure 1.1**: Scenario that led to the study problem

From Figure 1.1, the required solutions of object detection for both features and recognition can be concluded. Regarding the features; Features should be: Accurate, Fast, Invariant, and Easy to be computed. On the other hand; recognition
technique should be: Accurate, Fast, and Manage and reduce the effect of imbalance data on recognition quality.

1.3 Problem Statement

Full object detection system consists of three parts, features extraction, object recognition, and object localization. Most researches in literature focus only on one part of the whole system; each part affects the performance of the object detection system to some degree. Moreover, other challenges such as changes in rotation, scale, and imbalanced data effect are also affecting and complicating the system structure. Therefore, the performance of object detection system could be improved by introducing a new descriptor that produces fast, distinctive and invariant features, reducing the effect of imbalanced data by improving ensemble-based learning technique, and modifying the original interest points’ detector by which the system can be carried out faster.

1.4 Research Questions

Research questions are:

(i) How to overcome the shortages of existing descriptors in terms of accuracy and time consumption.

(ii) How to tackle the imbalance data and gain a high quality of recognition.
(iii) How to further reduce the detection errors; False Positive Rate (FPR) and False Positive (FP), and keep enough interest points for object localization.

(iv) How to speed up the whole system of object detection without impact on precision of detection.

1.5 Research Aim

The aim of this research is to introduce a general model for object detection in which it can overcome the limitation of existing methods regarding the descriptor, classification, and speed. The proposed model consists of three parts; first, new descriptor based on interest point detectors which is fast and accurate in the same time; second, an ensemble-based improved learner based on ANN technique applied on different environments; third, an improved scheme to speed up the object detection system; all are integrated in one model.

1.6 Research Objectives

The objectives of this research can be summarized as follows:

(i) To propose a new local descriptor that is faster and more accurate than the most common existing descriptors taken into account the stability under different conditions such as changes in scale and rotation.
To develop an effective supervised learning technique to learn the features of the proposed descriptor in (i) based on an ensemble learning in which it can increase the accuracy, reduce the error, and tackle the imbalance data challenge.

To propose a developed scheme that speed up the whole system resulting of the complexity in (ii) with no cost in quality of recognition.

1.7 Research Scope and Assumptions

Study scope and assumptions are as follows:

(i) Objects used in this study are; human faces and cars.

(ii) Scenes and environments in which the object is detected are:

a. Indoor and outdoor for faces detection.

b. Street and open space for cars detection.

(iii) Datasets used are:

a. Caltech 101 “faces”; for learning and testing.

b. PASCAL VOC 2009 “cars”; for testing.

c. SUN397; for learning.

d. Data for cars learning was collected by the researcher.

(iv) Measurements used in this research are: Correct classification Rate (CCR), True Positive Rate (TPR) or Detection Rate (DR), False Positive Rate (FPR), Precision, and False Positive (FP) or 1-Precision. The detail in chapter 3.
(v) Comparisons done as follows:

a. Comparing the proposed descriptor against SIFT and SURF descriptors based on measurements mentioned in (iv); besides, comparing the invariance to rotation and scale changes.

b. Comparing the ensemble-based classifiers against single-based classifier.

c. Time consumption comparisons; the proposed descriptor against SIFT and SURF descriptor, and the proposed scheme (that speed up the object detection system) against ensemble-based classifiers.

(vi) T-test was used to validate the obtained results regarding the FP error.

1.8 Research Significance

Object detection and its application play an important role in different fields as mentioned earlier; thus, this research takes the same importance beside the additional characteristics as follows:

(i) Usability: system can be used to detect any object by only providing positive and negative examples into the system; no extra changes are needed.

(ii) Development: different types of points’ detectors can be used if there, the structure of the model remains unchanged. In addition, easy to change the classifier used by another or adding another one beside the existing one, model structure remains also unchanged.
1.9 Summary of Research Contributions

Contributions of this study can be summarized as follows:

(i) A new local descriptor that is fast, accurate and overcome the existing descriptors in terms of speed, accuracy, and invariance.

(ii) A developed learning technique based on ensemble learning that produced higher accuracy than single-based learning technique and fixed the imbalance data problem.

(iii) A developed scheme that speeded up the object detection system by modifying the original SURF point’s detector.

(iv) Tow labelled datasets prepared from datasets mentioned in section 1.7 for training and testing phases.

1.10 Terms Definition

Due to using some terms in different domains, misconception or confusion may occurs; thus, this section provides the definition of terms used along this study, these definitions were taken from literature detailed in chapter 2.

(i) Interest Point

The most important point that can repeatable over scale-spaces, it can be edge or corner; sometimes it called key-point.

(ii) Detector
It is the method that detects the interest point.

(iii) **Descriptor**

Descriptor is the method that describe the detected interest point; describing the interest point means extract the substantial information about that point. Sometimes the method called *extractor* and the extracted information called *feature*.

(iv) **Object Recognition**

Object recognition is this study means, distinguishing the patterns that belonging to the object class from that they are not; it answer the question “Is there object within scene or not?” This includes the matching technique and learning methods.

(v) **Object Localization**

Objection localization means if there an object exist within scene, determine the location of that object; so, it answer the question “Where the object location within scene?”

(vi) **Object detection**

The meaning of object detection embodies both object recognition and localization altogether.

(vii) **Single-based classifier**

Normally, single-based classifier means only one classifier used for learning and classification of patterns; sometimes it called traditional or conventional method.

(viii) **Ensemble-based classifiers**

Ensemble-based methods often refer to a collection of classifiers, either by combining different classification techniques or variants of the same technique.
An imbalanced data

Data that has minority and majority among its classes is called imbalanced data; that’s mean the distribution of instances for classes is not equal.

1.11 Thesis Outline

This thesis consists of 7 chapters organized as follows: Chapter 1 introduces an overview on object detection system and its applications; besides, still problems and challenges, problem statement, study questions, study objectives, scope of the study, the aim and significance of the study and study contributions, all were explained. Chapter 2 illustrates the literature review and related studies to object detection, this includes the challenges with respect to features, recognition, localization, datasets, and evaluation measurements used in this area; in addition, the proposed solution in literature for each and still challenges. The research methodology of this study is illustrated in Chapter 3. Chapter 4 describes the proposed descriptor in detail which then integrated with a traditional classification technique as one model. Chapter 5 presents the solution for imbalanced data problem by developing an effective technique based on ensemble classifiers. A developed scheme for speeding up the object detection system is introduced in Chapter 6. Research summary, contributions, and the suggestions for the future work are detailed in Chapter 7.
REFERENCES


Li, H., Xu, T., Li, J., and Zhang, L. (2013). Face Recognition based on Improved SURF. Third International Conference on Intelligent System Design and Engineering Applications (ISDEA), Hong Kong, 755-758.


