REDEFINING HISTOGRAMS OF ORIENTED GRADIENTS DESCRIPTORS FOR HANDLING OCCLUSION IN PEOPLE DETECTION

AHMAD SANI

UNIVERSITI TEKNOLOGI MALAYSIA
REDEFINING HISTOGRAMS OF ORIENTED GRADIENTS DESCRIPTORS FOR HANDLING OCCLUSION IN PEOPLE DETECTION

AHMAD SANI

A dissertation submitted in partial fulfilment of the requirements for the award of the degree of Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JULY 2014
Dedicated with love to those I care...
ACKNOWLEDGEMENT

In the name of Allah, the most beneficent and the most merciful; all praise be to Him, the sole lord of the universe; and peace and blessing be upon the Prophet Muhammad, the seal of all prophets.

In this limited space I would like to show my gratitude to all people around; the family for their endless support, and friends for standing beside me. My thanks go to my supervisor, Assoc. Prof. Dr. Mohd Shafry Mohd Rahim, for his patience; and to Universiti Teknologi Malaysia and the Faculty of Computing as well. I was honoured to be a student at this university and under Dr. Shafry’s supervision.

A very special thank you is due to my mother for everything, and to my sister, Asst. Prof. Dr. Alawiyah, as well for her special favour. God bless them all.
Object detection is a big challenge for researchers to address the issues that affect accurate detections. The Histogram of oriented gradients (HOG) descriptors have been used extensively for object detection on challenging conditions with good results. However, occlusion remains a well-known issue where some parts of an object are only visible. Conventional frameworks using HOG descriptors for detecting whole human body are prone to occluded people and thus causing errors on implementations running these frameworks. This issue was addressed in different approaches like hybrid feature extraction algorithms. Rather than following such approaches, the problem is addressed in this research by redefining HOG descriptors to work in a part object (body) detection framework that is powered by a finite-state machine. Experiments were conducted to test a new method of extracting HOG descriptors for parts of human body via slicing a window’s HOG descriptor into four. The part body detection framework utilises each part’s descriptor whereas three detected parts are sufficient to declare that the window has an assumed instance. Support vector machines (SVM) were used to classify the extracted parts and a finite-state machine was employed for handling the detected parts. Training and testing were made on subsets from the INRIA Person Dataset and the results favoured the part body detection framework over the conventional whole body framework. For a test set of 50 positive images of occluded people, the part body detection framework successfully detected 46 true positive while the whole body framework detected 36. Moreover, the former had less false positive detection than the latter having 80 false positive windows comparing to 289.
ABSTRAK

Pengenalpastian objek, merupakan cabaran besar kepada para pengkaji bagi mengenalpasti isu-isu yang mempengaruhi ketepatan pengesanan. Pengesan Histogram of Oriented Gradients (HOG) telah digunakan secara meluas dalam mengenalpasti objek, dalam keadaan yang mencabar; dengan keputusan yang baik. Namun, pertindihan kekal sebagai isu yang sedia dimaklumi di mana hanya sesetengah bahagian sesuatu objek kelihatan. Rangka kerja menggunakan pengesan HOG untuk mengesan seluruh badan manusia cenderung menghalang keterlihatan manusia; yang menimbulkan ralat dalam pelaksanaannya. Isu ini diatasi melalui pendekatan berbeza seperti menggunakan algoritma yang mengesan feature campuran. Berbanding menggunakan pendekatan tersebut, masalah dirungkai dalam kajian ini melalui penelitian semula terhadap pengesan HOG supaya beroperasi mengikut rangka kerja pengesan bahagian objek(badan) yang dikuasai oleh finite-state machine. Kajian dijalankan bagi menguji kaedah baru untuk merungkai pengesan HOG mengikut bahagian badan manusia melalui penghirisan tetingkap pengesan HOG kepada empat(4) bahagian. Rangka kerja pengesan bahagian badan merungkai setiap bahagian; walaupun tiga(3) bahagian yang dikesan sudah mencukupi untuk mengesahkan tetingkap mempunyai imej seperti yang diandaikan. Support Vector Machines (SVM) berfungsi mengelaskan bahagian yang dikesan dan Finite-State Machine digunakan untuk mengendalikan bahagian yang dikesan. Latihan dan ujian dilakukan ke atas subset INRIA Person Dataset dan keputusan memihak kepada rangka kerja pengesan bahagian badan berbanding seluruh badan. Untuk ujian ke atas 50 set imej positif manusia yang terlindung, rangka kerja pengesan bahagian badan berjaya mengesan 46 imej positif benar manakala hanya 36 bagi rangka kerja seluruh badan. Malahan, rangka kerja pengesan bahagian badan mempunyai kadar pengesan positif palsu yang lebih rendah berbanding seluruh badan dengan perbandingan 80 dan 289 tetingkap positif palsu.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DECLARATION</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>DEDICATION</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td>ACKNOWLEDGEMENT</td>
<td>IV</td>
</tr>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>V</td>
</tr>
<tr>
<td></td>
<td>ABSTRAK</td>
<td>VI</td>
</tr>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>VII</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>XI</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>XIV</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Research Background</td>
<td>2</td>
</tr>
<tr>
<td>1.3</td>
<td>Problem Statement</td>
<td>4</td>
</tr>
<tr>
<td>1.4</td>
<td>Research Aim</td>
<td>5</td>
</tr>
<tr>
<td>1.5</td>
<td>Research Objective</td>
<td>5</td>
</tr>
<tr>
<td>1.6</td>
<td>Research Scope</td>
<td>5</td>
</tr>
<tr>
<td>1.7</td>
<td>Research Significance</td>
<td>6</td>
</tr>
<tr>
<td>1.8</td>
<td>Chapter Summary</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Feature Extraction</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.2.1 Haar wavelet Features</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.2.2 Scale Invariant Feature Transform</td>
<td>10</td>
</tr>
</tbody>
</table>
2.2.3 Histogram of Oriented Gradient 12

2.3 Object Detection Frameworks 15

2.3.1 Properties of Sliding Window Detectors 15

2.3.2 Handling Detected Parts in Part Object Detectors 16

2.3.3 Pre-processing in Object Detection Frameworks 19

2.4 Occlusion in Object Detection 20

2.4.1 Deformable Part Model 20

2.4.2 Hybrid Features with Local Binary Pattern 20

2.5 Object Detection Implementations 21

2.5.1 Road and Traffic Safety 21

2.5.2 Digital Camera and Robotics Vision 21

2.5.3 Entertainments 22

2.5.4 Contents Indexing 23

2.5.5 Surveillance 23

2.6 Classification 24

2.6.1 Neural Networks 24

2.6.2 Boosting 25

2.6.3 Support Vector Machine 26

2.7 Benchmarking and Evaluation 27

2.8 Summary for Related Works in Object Detection 30

2.9 Chapter Summary 36

3 METHODOLOGY

3.1 Introduction 37

3.2 Overview on the Research Methodology Framework 37

3.3 Dataset 38

3.4 Developing the Algorithm for Extracting the Redefined Descriptors 40
3.4.1 The Histograms of Oriented Gradients Descriptor Algorithm 40
3.4.2 Redefining HOG Descriptor for the Body Parts 42

3.5 Developing Detection Frameworks 44
3.5.1 Utilising the Whole Body Descriptor Only 45
3.5.2 Utilising Part Body Descriptors Only 45
3.5.3 Handling Detected Parts 46
3.5.4 Scanning and Grouping Multi-Detection 47

3.6 Classifier 48
3.7 Training the Detection Framework 48
3.8 Testing the Detection Framework 49
3.9 Performance Evaluation 49
3.10 Chapter Summary 50

4 RESULTS AND DISCUSSION
4.1 Introduction 51
4.2 The Environment Variables 52
4.3 Preparing the Dataset 53
4.4 Experiment Results Preview 53
4.5 Whole Body Detection Framework 55
  4.5.1 Training 56
  4.5.2 Testing 67
  4.5.3 Discussion 87
4.6 Part Body Detection Framework 88
  4.6.1 Training 89
  4.6.2 Choosing the Best Slice Size for the Redefined Part’s Descriptor 109
  4.6.3 Testing 111
  4.6.4 Discussion 132
4.7 Chapter Summary 133
5 CONCLUSION

5.1 Introduction 134
5.2 The Experiments Findings 134
  5.2.1 On Detection Frameworks 134
  5.2.2 On the Dataset, Training, and Testing 135
  5.2.3 On Slicing the Descriptor 135
5.3 Research Contribution 136
5.4 The Research Limitations and Future Works 137

REFERENCES 138

A THE DATASET

A.1 The Positive Train Set 147
A.2 The Negative Train Set 158
A.3 The Positive Test Set 159
A.4 The Negative Test Set 161
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Three Basic Haar Wavelets Rectangle Features (Oren et al., 1997)</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>Computing Integral Image; the Point at ((x,y)) is the Sum of All Pixels above and on Left (Viola and Jones, 2004)</td>
<td>9</td>
</tr>
<tr>
<td>2.3</td>
<td>Sample Frames from Videos where detected Human are bounded with Yellow Bounding Boxes (Viola et al., 2003)</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>Computing the Difference of Gaussian (DoG); an Image is convolved with Gaussians producing the Images on Left where They are subtracted producing the Images on Right (Lowe, 2004b)</td>
<td>11</td>
</tr>
<tr>
<td>2.5</td>
<td>HOG Performance; HOG Descriptors achieved Less Miss Rates comparing to Other State-of-the-Art Features in the MIT (Left) and INRIA (Right) Datasets (Dalal and Triggs, 2005)</td>
<td>13</td>
</tr>
<tr>
<td>2.6</td>
<td>Dividing detecting Window for Part Detection (Linh et al., 2011)</td>
<td>14</td>
</tr>
<tr>
<td>2.7</td>
<td>Experiment for sliding Detection Windows on an Image; Windows (as Green Bounding Boxes for detected Instances) overlap horizontally and vertically during the Exhaustive Search over the Image</td>
<td>16</td>
</tr>
</tbody>
</table>
2.8 A Part Object Detector examines Four Parts and merges Them into One Whole Object Detection (Mohan et al., 2001) 18

2.9 Classifiers in Cascade of Rejecters (Viola and Jones, 2004) 26

2.10 An Illustration of a Model with Two Types of Data classified in SVM; the Kernel is computed using \( Dx = wx + b = 0 \); the Margin (M*) is defined with the Nearest Two Support Vectors to the Kernel (Boser et al., 1992) 27

2.11 Samples from Daimler Pedestrian Benchmark. Images are Monochrome and relatively Lower in Resolution (Munder and Gavrila, 2006) 28

2.12 Samples from the MIT Pedestrian Dataset; the Background is Simple and People only are in Front and Rear Views 29

3.1 Overview on the Research Methodology Framework 39

3.2 Computing and classifying HOG Descriptors (Dalal and Triggs, 2005) 40

3.3 Overlapping Blocks coordinating with their Whole Body Window. During computing the Descriptor, the Input Window in (a) is divided into a Grid of Cells (8 × 16). A Block groups Four Cells (2 × 2) of Oriented Gradients Histograms, as in (c), and applies Normalisation then shifts by One Cell; resulting in Overlapping Blocks in (b). 42

3.4 The slicing Algorithm for extracting Part Body Descriptors 43

3.5 Mapping Table for the starting Indices of the overlapped Blocks in an HOG Descriptor 44

3.6 Demonstration for the detection Framework utilising Part Body Descriptors; (a) The
Framework declares a detected Person from Three detected Parts, unlike in (b) where Only Two Parts were detected. Sizes of Parts do not exactly reflect Those used in Experiments.

3.7 The Finite-State Machine for handling Part Body Detections

3.8 Invalid and Valid Samples from the INRIA Train Set; the Four Images on Left (a) are excluded from Training the SVMs as Each Body’s Parts do not fit well in the their designated Slices, while the Parts in the Other Four Images (b) do

4.1 Comparing the Two Detection Frameworks Performance on the Positive Test Set by counting the Images that were detected correctly (True Positive Case); Each Pair of Bars in the Chart shows the Results for testing with Different C Coefficients

4.2 Comparing the Two Detection Frameworks Performance on the Negative Test Set by counting the Windows that were detected incorrectly (False Positive Case). Each Doughnut in the Chart shows the Results for testing with Different C Coefficients

4.3 Slices on the 2-D Representation (Mapping Table) for a Whole Body’s Descriptor; Each part has Two Candidates where the nominated One (According to the Experiments in Table 4.7) is outlined with Dashed Lines
## 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>Highlights on Related Works</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Specifications for the Computer used for Experiments</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>The Experiment’s Parameters</td>
<td>52</td>
</tr>
<tr>
<td>4.3</td>
<td>The Performance of Each Detection Framework; the Numbers in Brackets indicate the Number of Images in a Test Set</td>
<td>54</td>
</tr>
<tr>
<td>4.4</td>
<td>Results from selecting the Highest Magnitudes (per Element) of Three Channels and extracting their HOG Descriptors for 20 Sample Windows (10 Positive Windows and 10 Negative Windows). The Visualised HOG Descriptors show Blocks in non-overlapping Forms</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Results from selecting the Highest Magnitudes of Three Channels and extracting their HOG Descriptors for 20 Sample Windows (10 Positive Windows and 10 Negative Windows). The Visualised HOG Descriptors show Blocks in non-overlapping Forms. Each SVM’s C resulted in its own Decision Function</td>
<td>68</td>
</tr>
<tr>
<td>4.6</td>
<td>Results from selecting the Highest Magnitudes (per Element) of Three Channels and extracting their HOG Descriptors for 20 Windows (10 Positive Windows and 10 Negative Windows).</td>
<td></td>
</tr>
</tbody>
</table>
The Visualised HOG Descriptors shows Blocks in non-overlapping Forms

4.7 Experiments Results for selecting the Best Slice Size

4.8 State Transition Table; the Four Columns from Right show the Input for the Machine and its Next States

4.9 Results from selecting the Highest Magnitudes of Three Channels and extracting their HOG Descriptors for 20 Windows (10 Positive Windows and 10 Negative Windows). The Visualised HOG Descriptors show Blocks in a non-overlapping Form. Each SVM’s $C$ resulted in its own Decision Function
CHAPTER 1

INTRODUCTION

1.1 Introduction

Detecting people in computer vision has been an active research area for its promising applications. The research in computer vision (and computer science in general) has given computers abilities inspired by nature such as the sense of sight; and nowadays there are many implementations in road-safety and surveillance having these abilities. Many algorithms that run these implementations are not exclusive only for people detection; researchers have extended them to other classes like animals, cars, and generally other objects. Detection frameworks or detectors vary from using pattern recognition algorithms in a holistic approach to exhaustive scanning at every location in the image. However, from observing nowadays detectors, there are made of two main processes of feature extraction and machine learning. Features are used as a more convenient way to describe object, thus they are often called descriptors, while machine learning is used for classifying these features to answer the question whether an instance of an object exist or not in an image. Designing detectors following this way of using two processes gives better understanding of problems in computer vision and thus they are addressed properly. The problem of occlusion, for instance, is well known to affect detections where parts of instances are overlapping or hidden by other objects; thus the research community proposed detecting any visible part in an approach known by part object detection (or part-based detection) rather than the other approach of whole object detection (the conventional way to detect the whole object). Other problems such as
view variations, poses, carrying belongings, clothes, scale, etc. were addressed as well as many features were proposed to handle such problems. However, the histograms of oriented gradients (HOG) descriptors proved to be robust to many issues and this dissertation aims to improve the use of these descriptors via redefining them into a framework of part object detection.

1.2 Research Background

Object detection is a task with different issues regarding how the objects appear in images. First, there was the question of how to find/detect instances using their shapes (Borgefors, 1988); object were matched to a template that is rigid to any object variations, thus more convenient ways to describe object were required. Oren et al. (1997) proposed a detection framework that has influenced many nowadays detectors; they proposed an exhaustive approach for scanning images for any instance using sliding windows where Haar wavelet features are extracted then a support vector machine (machine learning) classifies whether instances are there or not. Viola and Jones (2001) introduced their face detection (a specific domain in object detection) framework and went successful for real-time performance thanks to the fast computing in Haar wavelets and the integral image.

However, while these features have good detection performance when the variation of intensities between the object and the background is high, the detection frameworks faces troubles when the intensities variation are low due to issues related to cluttered background, clothes, belongings, etc. where texture-based features are not suitable. Gradient-based features came to be better for such situations where an object’s silhouette can be captured well. Adding the statistical concepts of orientation histograms at local patches had given more robust features such as the scale-invariant feature transform (SIFT) by Lowe (1999). SIFT returns keypoints used basically for recognising objects, yet Mikolajczyk et al. (2004) implemented SIFT features for object detection where their detector performed well. This performance motivated
further studying local histograms of oriented gradients, in terms of optimising computation time (Yan and Sukthankar, 2004) or increasing detection accuracy (Mikolajczyk and Schmid, 2005). Yet among many, the histogram of oriented gradients (HOG) descriptor (Dalal and Triggs, 2005) managed to capture the human body silhouette efficiently where experiments on HOG descriptors have showed excellent performance on the MIT pedestrian dataset and the INRIA person dataset. HOG descriptors work well in capturing the whole human body shape that may appear in different poses, views, clothes, and under different lightening conditions (the illumination factor). Yet, like SIFT, it has slower computation time comparing to fast feature extraction algorithms like Haar wavelets. There have been some algorithms for HOG-based descriptors aimed for speeding up the process by selecting likely valid features using machine learning methods, for example using AdaBoost for feature selection (Qiang et al., 2006). While others aimed for increasing the accuracy and detection rate regardless of time costs (as computer hardware are advancing), such as a work by Satpathy et al. (2014).

In general, the amount of works increased for people detection (as well as object detection) research in the last decade; a survey by Geronimo et al. (2010) discussed the domain of people detection becoming a very active topic and draw the attention for building pedestrian protection systems (PPS) in advanced driver assistance systems (ADAS). Each work aimed to addressing challenges in object detection. But since HOG descriptors has done well with many issues, they were used widely for human and object detection in general (Kittipanya-ngam and Lung, 2011). Moreover, HOG descriptors are favoured over SIFT since the former does not require a license while the SIFT algorithm was patented (Lowe, 2004a) and requires fees for its implementations. However, the problem of occlusion remains a great challenge in computer vision to be covered and researchers followed different approaches to handle it. However, part object detection frameworks remain a naïve yet effective way to detect any visible part and join them using a model. Based on these frameworks, researchers introduced large but well-developed detectors employing a number of algorithms. For example, a work by Felzenszwalb et al. (2010) presented a part object detection framework that uses deformable part model to classify a number of objects and was able to handle simple occlusion. Wang et al.
presented a hybrid detection framework for handling occluded scenes using HOG descriptors and local binary pattern (LBP) features where it aims first to find occluded regions and then run the part detector where needed. This work was followed recently in similar concept by Marin et al. (2014). Although all these works handled varying levels of occlusion well, the framework inside remains complex with many algorithm that cost time.

Using simpler frameworks may lead to affordable answers for handling occlusion rather than consume more time on e.g. extracting different descriptors; what if the same descriptor is used redefined in a part object detection framework? And as a part object detector, what model will be used to join the detected parts?

1.3 Problem Statement

Although HOG descriptors have been utilised extensively for human and objects detection in general, their performance drops down when they deal with occlusion. Occlusion is a problem where parts of people are hidden by others causing more missing rates while detecting. In the available works that were reviewed, occlusion was handled in different detection frameworks using different algorithms and/or features. However, since part object detection frameworks can be used to handle occlusion and HOG has better performance on separate objects, occlusion can be handled when HOG is redefined to work in a part object detection framework that is less complex (without the need to use features other than the available HOG descriptor). This framework adds for HOG descriptors the robustness to occlusion and introduces a simple model that handles the detected part efficiently.
1.4 Research Aim

This dissertation aims to develop a variant HOG descriptor that is capable to work in a part object detection framework that is robust to occlusion. The framework should be less complex by redefining and utilising the available HOG descriptors and properly handles the detected parts.

1.5 Research Objective

From the discussion in the previous sections, the objectives for this dissertation are:

1. to redefine HOG descriptors for parts of the human body in a part object detector;
2. to develop a part object detection framework able to handle occluded people; and
3. to evaluate the performance of the new HOG descriptors and the developed detector.

1.6 Research Scope

This dissertation with its objectives is outlined to the following:

1. Input images used are still images, whether they were originated from photographs or still frames (snapshots) from videos.
2. Input images are pre-processed by balancing the illumination.
3. The detection framework is made for detecting people walking or standing; no consideration for people shown in half such as sitting on benches or on bicycles and so forth.
4. The occlusion is defined by one part of the human body, either one of the arms or of the legs, being hidden by another.

5. People in input images may wear any clothes, carry belongings, and be in different poses or from different views as long as there are visible three parts to detect.

6. The detection runs on one scale of the image and input images for testing are pre-processed by cropping them to make all people be in one defined size.

1.7 Research Significance

Object detection is a promising research area with endless implementations such as ADAS vehicle systems, robotics, and surveillance. However, more challenges occur as implementations grow. Occlusion is an everyday situation that affects any system relying on object detection frameworks. Therefore, there is a need to further the research on new HOG-based descriptors which are robust to occlusion and less complex to implement. Furthermore, the dissertation’s detection framework that utilises the new descriptors can be extended to work on other object classes for general object detection. The dissertation will contribute to different research areas, i.e. occlusion in object detection, introducing redefined HOG descriptors for human parts, and a new part object detection frameworks.

1.8 Chapter Summary

This chapter discussed the dissertation topic and the problem it addresses. In a nutshell, a redefined HOG descriptor are used in a part object framework for detecting people whose parts are hidden by other objects. Furthering such research will help overcoming the occlusion problem that affects performance in object detection. The next chapter discusses some works related to the dissertation topic.
REFERENCES


