Detection of Texture-less Occluded Objects by Deformable Part Modals

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of

Master of Applied Science

in

Electronic Systems Engineering

University of Regina

By

Faisal Iqbal

Regina, Saskatchewan

January, 2017

Copyright 2016: F. Iqbal
Faisal Iqbal, candidate for the degree of Master of Applied Science in Electronic Systems Engineering, has presented a thesis titled, *Detection of Textureless Occluded Objects by Deformable Part Models*, in an oral examination held on December 7, 2016. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner: Dr. Xue-Dong Yang, Department of Computer Science

Supervisor: Dr. Raman Paranjape, Electronic Systems Engineering

Committee Member: Dr. Paul Laforge, Electronic Systems Engineering

Committee Member: Dr. Craig Gelowitz, Software Systems Engineering

Chair of Defense: Dr. Fanhua Zeng, Petroleum Systems Engineering
Abstract

Detection of objects in images has been a popular area of research in computer vision. Researchers have been working on the automatic detection of both fully visible and occluded objects. In this thesis work, detection of occluded objects is further examined. This work targets occluded objects which are simple in shape and do not contain complex textures. Texture-rich objects are, in general, more easily detected; but occlusion of texture-less objects is a more challenging task. In this work we focus on the texture-free occluded images of mugs and cups. This thesis examines the effectiveness of part-based detection technique for the detection of these texture-less occluded objects. We believe that this approach is more effective than traditional approaches that look for the color and edges/shape of the objects only. The difficulty arises because the extent of occlusion is not uniform and occlusion is usually caused by cluttered environments creating shadows and many unpredictable edges. These issues decrease the performance of edge and color based detectors to a great extent. Our approach uses state-of-the-art deformable part models [11] that are applied to histograms of oriented gradients [7] as features. Instead of considering occlusion in general terms, we quantitatively define the degree of occlusion in percentage and train a latent SVM formulation on a mixture of occlusion models. Due to the lack of availability of a standardized occlusion dataset, we have used the well-known ETHZ shape classes dataset [16] along with flickr images to create a synthetic dataset of occluded mugs and cups. The proposed methodology is compared with the shape-
based detector of [15] [14] by developing precision-recall curves for both methods. We see that our methodology consistently outperforms edge/shape-based technique by an average precision of 0.30. We therefore conclude that our approach warrants further investigation and may provide better detection of texture-free occluded objects.

**keywords:** Computer Vision, Machine Learning, object detection, Support Vector Machine, occlusion detection
Acknowledgements

I would like to express my gratitude to my adviser Dr. Raman B. Paranjape for his valuable guidance and encouragement throughout this research work. He regularly reviewed this research and suggested improvements. I am very thankful to Faculty of Graduate Studies and Research for their financial support during my studies.
# Table of Contents

Abstract i

Acknowledgments iii

Table of Contents vi

List of Tables vii

List of Figures ix

1 Introduction and Overview 1

1.1 Introduction ............................................. 1
1.2 Problem Statement ......................................... 7
1.3 Contributions ............................................... 8
1.4 Thesis Outline ............................................. 9

2 Development in Object Detection methods and Related Work 11

2.1 Development in Object Detection .......................... 11
2.1.1 Geometric Models ...................................... 12
2.1.2 Brightness based Methods .............................. 13
2.1.3 Feature Based Methods ................................. 13
2.1.3.1 Interest Point detectors ............................ 14
2.1.3.2 Scale Invariant Feature Transform .................. 14
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.3.3 Speeded Up Robust Features</td>
<td>15</td>
</tr>
<tr>
<td>2.1.3.4 Histogram of Oriented Gradients</td>
<td>16</td>
</tr>
<tr>
<td>2.1.3.5 Deformable Part Models</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Occlusion handling techniques for Human and Cars</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Methods for detection of Texture-less Objects</td>
<td>20</td>
</tr>
<tr>
<td>3 Methodology</td>
<td>24</td>
</tr>
<tr>
<td>3.1 Concept of Image Convolution, Correlation and inner product</td>
<td>25</td>
</tr>
<tr>
<td>3.2 Histogram of Oriented Gradients</td>
<td>27</td>
</tr>
<tr>
<td>3.3 Deformable Part Models</td>
<td>29</td>
</tr>
<tr>
<td>3.3.1 Pyramid Representation</td>
<td>30</td>
</tr>
<tr>
<td>3.3.2 Defining Models</td>
<td>31</td>
</tr>
<tr>
<td>3.3.3 Matching</td>
<td>33</td>
</tr>
<tr>
<td>3.4 Machine learning</td>
<td>33</td>
</tr>
<tr>
<td>3.5 Optimization of Objective Function</td>
<td>35</td>
</tr>
<tr>
<td>3.5.1 Defining Latent Values</td>
<td>36</td>
</tr>
<tr>
<td>3.5.2 ComputationS of $\beta$</td>
<td>36</td>
</tr>
<tr>
<td>3.6 Data mining of training data</td>
<td>37</td>
</tr>
<tr>
<td>3.6.1 Data mining in Classical SVM</td>
<td>38</td>
</tr>
<tr>
<td>3.6.2 Data mining in Latent SVM</td>
<td>39</td>
</tr>
<tr>
<td>3.7 Model Training</td>
<td>41</td>
</tr>
<tr>
<td>3.7.1 Preparation of Training data</td>
<td>41</td>
</tr>
<tr>
<td>3.7.2 Defining latent values</td>
<td>42</td>
</tr>
<tr>
<td>4 Experimental Results</td>
<td>43</td>
</tr>
<tr>
<td>4.1 Dataset</td>
<td>43</td>
</tr>
<tr>
<td>4.2 Test procedure</td>
<td>46</td>
</tr>
<tr>
<td>4.2.1 Precision-Recall</td>
<td>46</td>
</tr>
</tbody>
</table>
List of Tables

4.1  Average Precision of DPM with HoG features . . . . . . . . . . . . . 54
4.2  Average Precision of edge based object detection method . . . . . . 56
List of Figures

1.1 Occlusion happens almost everywhere in natural arrangement . . . . 2
1.2 Edge-only image on the left and HoG only image on the right . . . . 4
1.3 Overview of the complete methodology . . . . . . . . . . . . . . . . . 5

2.1 Polyhedral models for objects proposed by L.G.Roberts [13](Left) and
generalized cylinders by Thomas Binford [1] (Right) . . . . . . . . . . 12
2.2 Difference of Gaussian(Left) and 4x4 Region Descriptor for SIFT de-
scriptor [31] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15
2.3 Image of a person on the left and gradient weights learned by support
vector machine classifier [7] . . . . . . . . . . . . . . . . . . . . . . . 17
2.4 Image of a person on the left and part based HoG representation(Middle)
and allowed range for displacement of parts (Right) shown as white re-
gions [12] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

3.1 Representation of cells and blocks and 50% sliding blocks. Aggregation
bin are shown on the top right corner . . . . . . . . . . . . . . . . . . 28
3.2 Image Pyramid showing only half resolution than the next level . . . 30
3.3 Image and Feature pyramids are shown side by side. It can be seen
that part filters are applied at twice the resolution as that of root filter 32

4.1 Top row shows images with only 25% occlusion while middle row has
examples of 50% occlusion. Bottom row has 75% occlusion in images 46
4.2 Precision-Recall explanation pictorially represented 47
4.3 Concept of IOU and 50% IOU shown pictorially 47
4.4 PR curve for DPM based occluded texture-less object detection with prior knowledge of degree of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results 50
4.5 PR curve for DPM based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results 52
4.6 PR curve for DPM based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results 53
4.7 PR curve for edge based occluded texture-less object detection with prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion testset results 57
4.8 PR curve for edge based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion occlusion testset results 58
4.9 Images on the left show detection by deformable part models with red bounding box as overall detection window. Images on the right show false detection by edge based method on same images 59
**Abbreviations**

AP  Average Precision

CD  Compact Disk

DPM  Deformable Part Models

FT  Fourier Transform

GLOH  Gradient Location and Orientation Histogram

GPU  Graphics Processing Unit

HoG  Histogram of Oriented Gradients

IoU  Intersection over Union

PCA  Principal Component Analysis

SIFT  Scale Invariant Feature Transform

SURF  Speeded up Robust Features

SVM  Support Vector Machine

VTR  Visual Teach and Repeat

WT  Wavelet Transform
Chapter 1

Introduction and Overview

1.1 Introduction

Object detection has gained much interest from researchers in recent decades due to the use of high processing machines and GPUs. Object detection finds its application in multiple areas specially in the field of robotics where robots are being used to help people in their daily life, to carry out potentially harmful tasks, to carry out space expeditions, to do autonomous surveillance etc. The need for robust object detection techniques is increasing to improve the performance of autonomous systems. Many sophisticated techniques have been developed for object detection. In most of the recent techniques, features of objects are extracted from training images and classifiers are trained on these features by machine learning algorithms. However the performance of the classifier is badly affected when the objects get occluded.

Occlusion is a condition when an object is partially hidden by other objects. Occlusion happens almost everywhere in natural arrangement. Figure 1.1 shows images of two scenes that show occlusion by natural arrangement. It is not an elusive task to detect an occluded object by human brain however the machine is not robust to these changes. Sophisticated research is present to deal with occlusion of objects.
Figure 1.1: Occlusion happens almost every where in natural arrangement

but most of the work deals with objects in motion such as object tracking in video sequence [37][52]. Most of these methods use optical flow to distinguish between occluder and occludee objects. Use of multiple cameras to detect and localize object cameras [8][35] has also shown good results when object gets occluded. A problem with multiple cameras is that such an environment with multiple cameras at required positions is only available in controlled situations. There is, however, less exploration to deal with occlusion of static objects in case of single camera image. Algorithms based on optical flow does not support the occlusion when all objects are stationary. We can categorize objects in two major types i) Texture-rich objects and ii) Texture-less objects. A key point on texture rich content can easily be detected by interest point detectors such as Harris Corner detector [22], Shi-Tumasi [42] and SUSAN [46]. Interest point descriptors capture the characteristics of the neighborhood of the interest points. Scale Invariant Feature Transform (SIFT) [30], Speeded Up Robust Features (SURF) [3], Gradient Location and Orientation Histogram (GLOH) [34] and
Histogram of Oriented Gradients (HoG) [7] are commonly used interest point descriptors. However, interest point detectors do not perform well on texture-less objects because these objects are simple in their shapes and do not have complex curves and sharp corners. Also the aforementioned interest point detectors perform better on the same objects for whom the interest points were recorded. The examples include Visual Teach and Repeat (VTR) applications where an autonomous robot is supposed to traverse an already learnt path by finding the same interest points during path learning process.

For texture-less objects such as mugs, cups, compact disk (CD) etc. other features are commonly used such as edge maps [16] and Histogram of oriented gradients (HoG) descriptor [24] [25]. A classifier is usually trained on these features obtained from training images dataset. Edge based detection focuses on the shape outline of the objects and neglects the texture information around the edges of the object. Furthermore, when the object gets occluded, it loses information of its edges and the edges of occluding objects appear in the edge-only map. In contrast, histogram of Oriented gradient (HoG) features capture the edge information along with texture near edges and on the object itself. Figure 1.2 shows ‘edge only’ and ‘HoG magnitude map’ of an image. It can be observed that energy is concentrated on the edges in edge-only image while, in case of HoG, it is also distributed over the objects. Deformable Part models (DPMs) introduced by Felzenszwalb et al. [11] have shown remarkable success in object detection specially human detection. These models use HoGs as features. These models are used in this thesis work to obtain average precision of detection on the dataset of texture-less occluded objects. To obtain the DPM model for the occluded mugs, images are collected to make an image collection. Occluded dataset is created synthetically by three degrees of occlusion namely 25%, 50% and 75% occlusion. After the creation of the dataset, object annotations are prepared for
Figure 1.2: Edge-only image on the left and HoG only image on the right

each image in the dataset. Dataset is randomly divided into training, validation and testsets. Deformable part models are trained for three different occlusion cases and testing is done to get optimal detection solution for occluded texture-less objects. In the last step, results are compared with an edge based method trained and tested on same datasets and comparison is done. Overview of the training and testing is shown in the Figure 1.3. Benefits of Deformable part models with HoG features are briefly stated below:

• **Insensitivity of HoGs to brightness**: HoGs are intrinsically protected from additive brightness. The reason behind this immunity is the pixel difference taken in the calculation of gradients. The additive part is canceled in the difference.

• **Insensitivity of HoGs to contrast**: To compute HoG features, image is divided into blocks which are further consisted of cells. To compensate for local contrast effect, normalization of individual block is done. To further reduce the effect of local changes in the neighborhood of block, 50% overlapping blocks are traversed and normalization is done. At the end, the whole HoG descriptor is also normalized.

• **Object Detection at all scales**: The scale of an object is not typically known in a given single camera image. To deal with the scale problem, the image is
Figure 1.3: Overview of the complete methodology
smoothed and sub-sampled to create a scale space. HoGs are calculated on all levels of this scale pyramid of image to capture the object in a wide range of scale.

- **Deformable arrangement**: Objects belonging to a particular class are not exactly same. These differ visually in size, shape and color from each other. This is known as intra-class variation. In this situation, simple techniques such as template matching, is probably not a good solution. Apart from learning an overall texture information, DPMs also learn about different parts of the object. An object is searched for parts and these are allowed to be displaced from their mean position to some extent making the object detectable even though the parts are not at their ideal position. This makes DPM a good choice for objects with intra class variation.

- **Support Vector Machine as Classifier**: Support Vector Machine is a powerful tool for classification. Feature descriptors that are only trained on positive examples [54] usually do not perform well on negative images. In contrast SVM is a supervised learning algorithm trained on a set of positive and negative examples with maximum margin boundary between them. DPMs are trained on SVM for better classification results.

- **Training with partially labeled data**: Latent SVM (LSVM) formulation is used to train DPMs which use partially labeled data for training. A bounding box is drawn on the object without labels of the parts. The Algorithm searches the parts itself based on maximum energy in gradients domain.
1.2 Problem Statement

The objective of this thesis is to deal with occlusion of texture-less objects with Deformable Part Models [11] [12]. The class of the object selected for this task is a coffee mug, motivated by the problem of object detection in kitchen clutter. Its applications include robots helping human in their homes with cluttered environment, robots performing rescue operations etc. The problem can be divided into three phases, collection of the dataset, training of occlusion models and testing.

- Due to unavailability of publicly available dataset of mugs, images are collected from online resources. Synthetic datasets are created for training, validation and testing purpose. The size of image is one of major role in computation complexity of the algorithm. The size of images for this research work is chosen based on well known dataset [9]. Medium size images are chosen with dimension around 640x425 (width x height). Further detail is given in the next section.

- Instead of considering occlusion in a general sense, three different extents of occlusion are taken into consideration namely 25%, 50% and 75%. Deformable part models are trained using the synthetic dataset mentioned above. Detailed testing is performed for detection of occluded objects with these models and to suggest optimal solution.

- Testing is performed by the trained model in a sequence of informed-degree-of-occlusion toward uninformed-occlusion to observe the results. Bounding box localization is used for detection of occluded object. Intersection over union (IoU) is used to classify a detection as true or false. More than 50% IoU of bounding box with ground-truth is selected as true detection sometimes more lenient IOU are used by researchers such as 20% considered in [15]. To perform testing, a synthetic dataset of occluded images is created. Precision-Recall curves are used to see performance of detection. Detailed testing is performed
to get an optimal output of the system. At the very first stage, each model is applied on dataset of corresponding degree of occlusion. After that, each model is applied on a mixture of occlusion datasets without prior information about occlusion. In the third stage, a mixture of occlusion models is applied on a mixture of occlusion dataset with rule based reasoning for occlusion detection.

This research project consists of image collection, synthetic dataset and generation of annotation, training of DPMs on varying occlusion extent, testing these models on three testsets of occluded object and suggesting the optimal solution for detection of texture-less occluded objects of Mugs and Cups. Localization of an instance is marked by an enclosing rectangle around the detected region of the image. Condition for true detections are defined in terms of overlap between ground-truth rectangle and detected rectangle. In this thesis intersection over union (IoU) criterion is used and evaluation is done using IoU of more than 50

1.3 Contributions

- Successful implementation of DPM on texture-less occluded objects detection. Three models are trained on training images of texture-less objects with three different extent of occlusion. Testing is performed with these models on synthetically created dataset. Precision-Recall curves have been used to present results of detection. Average precision (AP) is listed in tabular form to show and compare the detection results.

- Results, obtained during the testing of object detection on test images, that the models trained on 25% and 75% occlusion perform badly on mixture of data set while model of 50% occlusion works best with an average precision of almost same as that of individual detection.

- Another contribution is made by making a dataset publically available. The
only data set of mugs available publicly is ETHZ shape classes data set [14] that contain 48 images of mugs. A dataset of 1000 images is collected and made available online. These images are collected from online image source ‘flickr’. Well known PASCAL Visual Object Classes project has been using images from flickr. Details of each image is also provided for tracking and acknowledgement of the photographer.

• Three test sets of occluded mugs with 25%, 50% and 75% occlusion of mugs are created synthetically. Each of these datasets contain 200 synthetic created images. Detection results are calculated on these test sets in this thesis. These test sets are also available on line.

1.4 Thesis Outline

This thesis is written in chapter form in the following arrangement:

• Chapter 1: Current research trends in the field of object detection are discussed along with some applications. Advantages of DPMs are discussed for occluded objects detection. Problem statement is defined and contribution are listed.

• Chapter 2: Explains the development in the field of object detection. An overview of Interest point detectors and descriptors is presented along with brief explanation. Benefits of HoGs and their use in DPM is mentioned.

• Chapter 3: Methodology is explained in this chapter. Structure of HoG features is explained. The concept of image and feature pyramids are explained and shown pictorially. Score of a model is defined in mathematical equation that combine the score of root and part filters. Support Vector machine’s objective function defined and training procedure is explained.
• Chapter 4: Detailed results of DPM are presented in this chapter. Response of
different models with prior knowledge of degree of occlusion and without any
prior information is explained. Optimal solution is suggested based on results.

• Chapter 5: Conclusion of this research work is presented along with future work.
Chapter 2

Development in Object Detection
methods and Related Work

In this chapter development of object detection methods is presented, and along-with that, related research work is described that utilized these methods in their original or modified forms. This thesis work consists of detection of texture-less objects with occlusion. To the best of my knowledge, much less work has been done related to occlusion handling of texture-less objects. As such, research work related to occlusion handling for other objects such as people, car etc. is presented. Also detection techniques of texture-less objects without occlusion is presented.

2.1 Development in Object Detection

Since its beginning, researchers in computer vision, have been proposing different methods to model objects present in images. It is, however, not an easy task to model different kinds of objects by a single approach. Apart from a wide variety of object classes, intra class variation is also present. Generally there are three major approaches that have been used for object detection.

1. Geometric Models
2. Brightness based Methods

3. Feature based Methods

2.1.1 Geometric Models

Geometric Models are one of the very initial approaches to model objects and for almost six decades these models were very popular among researchers. In this approach, objects were modeled as predefined 3D polyhedral shapes and perspective projection of the object as polygons were detected by heuristic reasoning. L.G Robert [13] developed a detection system based on this approach. Moreover he developed his own edge detection algorithm to detect polygons of objects. There are number of intrinsic weaknesses associated with this algorithm. For example curved surfaces, complex shapes, transparent objects, inter and intra class variations of objects etc. In [1], Thomas Binford employed the concept of generalized cylinders to deal with curved shapes. Although the algorithm proposed by Thomas elicits better results when compared against traditional algorithms, it is still unable to deal with all type of shapes, their characteristics and deformations due to a large inter and intra-class variation of objects. Both models are shown in Figure 2.1

Figure 2.1: Polyhedral models for objects proposed by L.G.Roberts [13](Left) and generalized cylinders by Thomas Binford [1] (Right)
2.1.2 Brightness based Methods

These methods mainly rely on brightness information of the object present in an image. Generally two frameworks are used in this regard i) Template based methods ii) Feature based methods. In template based methods, matching of objects is done by finding correspondence between gray scale pixels of template and test image such as template matching. Authors of [53] suggested a deformable template which is based on peaks and valleys present in form of pixel intensity values of object. This method shows some robustness to deformation but it requires templates that are to be designed by hand so it is not scalable. The authors in [48] [6] proposed the similar idea but instead of directly comparing the gray pixels of test and template image, they warp images onto one another by correspondences learned during the training phase. Although template based methods provide naive and elegant solution for object detection, the performance of these methods is greatly affected by the curse of dimensionality. In the second category, pixel to pixel processing is not done. In order to deal with the dimensionality of data, features that represent the most important information in the data are generally extracted. PCA, FT and WT are few names to such methods that are used for dimensionality reduction as well as to extract key features from the data. The features extracted from the data are eventually processed by machine learning tools to learn patterns found in derived features. Many applications are found such as hand written character recognition [28], diagnosis of medical condition in Radiographic images by wavelet transform [39]. Methods mentioned in the first category are not commonly used however second category methods are still common and useful.

2.1.3 Feature Based Methods

Feature based methods, also known as appearance based methods [29], have been very popular in the last two decades. These methods try to capture shape properties and
interest points on objects such as shape silhouette, corners, edges of the objects, pixel gradients etc., instead of just modeling them as geometric polygons or matching pixel gray levels.

2.1.3.1 Interest Point detectors

Interest point detectors look for corner points that are characterized by change in intensity in every direction. Harris corner detector, Shi-Tomasi corner detector, SU-SAN corner detector and Difference of Gaussian (DoG) are commonly used. However these detectors are not robust against scale and shape variation of objects.

2.1.3.2 Scale Invariant Feature Transform

To overcome these weaknesses a well-acknowledged Interest point descriptor was presented by David G. Lowe [31] named as Scale Invariant Feature Transform (SIFT). SIFT is scale invariant as it looks for interest point over a range of image scale space obtained by generating an image pyramid by gaussian-kernel smoothing and subsampling of a given image. Lowe presented an alternate of Laplacian of Gaussian (LoG) as Difference of Gaussian (DoG) to look for zero crossings produced by second order derivative at the edges, corner points and boundaries in an image. To incorporate repeatability, a region of 16x16 pixels around interest points is described in form of histogram of gradients calculated in that region. Gradient directions are assigned to eight predefined bins by taking 4x4 set of pixels. This gives a descriptor of 4x4 with 8 orientation giving rise to a descriptor vector of 4x4x8=128 values for each interest point. Figure 2.2 show these concepts pictorially. In the learning phase, SIFT points are detected and stored in the memory and are used for matching and finding the same objects in test images.
This descriptor has shown robustness against scale changes, rotation, noise, stretching, contrast and brightness changes. One major drawback in the use of SIFT is its computation time because an image size of 512x512 gives an order of 1000 key points to be processed [30]. This makes it less applicable for real time applications.

2.1.3.3 Speeded Up Robust Features

To overcome the computation complexity, Speeded up Robust Features (SURF) was coined by Bay et al. [2]. They used the concept of integral images to compute sum of a rectangular region in an image. Three additions and four memory access operations are required to get the sum of a rectangular region in an image. Moreover they approximated second order Gaussian derivative with box filters with peaks as ones and valleys as zero. They used the integral images to get convolution result of these box filters. The computation time was reduced over the scale space because the size of filter was not affecting the computation time due to fixed number of operations involved in integral image concept. They further used Haar Wavelet response for 4x4 sub region around the interest point in the horizontal and vertical direction with origi-
inal and absolute values giving 4 response values. This gives a feature vector length of 
$4 \times 4 \times 4 = 64$. Similar to SIFT, key point descriptors learned during training are used 
to search the same object in test images. However, both SIFT and SURF descriptors 
are suitable for the applications where the testing is to be performed on the exactly 
same objects from which the key points were detected. Examples include simultane-
ous Localization and mapping (SLAM), 3D reconstruction, panoramic image stitching 
etc. They are not suitable, however, for detection of general category of objects such 
as pedestrian detection, pose estimation etc.

2.1.3.4 Histogram of Oriented Gradients

Gradient based region descriptors like SIFT [30] and SURF [2] have performed ex-
traordinarily. These descriptor are based on 2D gradient values calculated on each 
pixel with normalization to cancel the brightness and contrast changes in images. 
This concept had been used by researchers [18] [19] [33] but it reached its maturity 
when used as spatially significant point descriptors and performing normalization of 
these descriptors [7]. In [7], Dalal and Triggs used the similar idea and introduced 
Histogram of oriented gradients (HoGs) which significantly outperformed other ex-
isting [4] [49] human detection algorithms. Dalal and Triggs further generalized this 
concept of gradient histograms and trained a support vector machine classifier with 
HoG features for pedestrian detection. They obtained gradients on pedestrian im-
ages and divided the region into smaller parts called cells. Gradient directions were 
assigned to eight bins with linear interpolation of gradients to nearby bins. In order 
to make their final descriptor contrast invariant, cells were combined to make a block 
which was then normalized in 50% overlapping manners.
These blocks were concatenated together to make a feature vector which was fed to soft margin linear Support Vector Machine for training. Figure 2.3 shows a person and gradient weights learned by support vector machine classifier.

2.1.3.5 Deformable Part Models

Detection results presented by Dalal and Triggs were impressive in that they outperformed wavelet based methods of human detection [7]. This state of the art was further improved by Felzenszwalb et al. [12] by incorporating the concept of deformable part model concept which was introduced well before in 1973 by Fischler and Elschlager [17]. Deformable part models (DPM) are classic till today due to their ability to describe an object in form of its parts and to incorporate shape deformation in their formulation. DPM by Felzenszwalb et al. [12] use HoG features in their models. The training of DPMs require very lightly annotated data which is a bounding box around the object. DPM training is a hierarchical procedure in which an overall shape filter (named as root filter) is learnt by SVM in form of HoG
features and then a predefined number of parts are searched and incorporated into model. Parts are searched based on high energy regions in HoG feature domain. The location of each part is also defined with respect to an anchor point in its root filter. Figure 2.4 shows DPM for human. The concept of deformation introduced in [17] considers different parts of objects connected together via springs that allows them to be displaced from their respective positions to some extent but with some penalty for being displaced from their ideal location. This displacement cost could be selected as linear or quadratic depending upon application.

DPMs are selected for this thesis work to detect texture-less occluded objects of Mugs and cups as these objects are very common in domestic environment. It is rather difficult to detect occlusion in the case of texture-less objects primarily due to simplicity of the shape of these objects. Interest point descriptors (SIFT, SURF) perform well on the objects that are rich in corners and blobs and do not vary in shape. Texture-less objects, however, do not comply with these conditions. Now an overview of this work is presented along with the work dealing with occlusion of other objects such as pedestrians, cars etc.

Figure 2.4: Image of a person on the left and part based HoG representation (Middle) and allowed range for displacement of parts (Right) shown as white regions [12]
2.2 Occlusion handling techniques for Human and Cars

Researchers have made efforts to deal with the problem of occlusion for object tracking applications in videos such as video surveillance, control of traffic, medical imaging etc. [44] [45] [5]. These applications typically involve sequence of frames available by a video stream which makes manipulation of pixels easier as compared to static images. The use of contextual information and optical flow makes tracking of objects feasible even in case of occlusion. Authors in [44] proposed a method to track multiple persons in crowded environment by using part based models to detect people in the scene. They proposed part based models because the crowded environment often creates occlusions between people. In this case, single-detection-window based models fail as the occlusion is caused to the people by other people while part based model localize the visible portion and vote for the detection as the information of object is passed between frames to track the object. Latent SVM is used to train part based human models.

Similar to this approach, [51] used HoG features to detect humans in case of occlusion. An overall detection window is used to search for fully visible object while a detection window with occlusion is searched with part filters to detect occlusion. To detect a window with occlusion, authors have utilized a unique observation. They observed that the inner product between the model weight and the feature vector results in negative value when any window contains object with occlusion. Filter responses are obtained by a sliding window approach and the response of the inner product is searched for negative values. Those portions with negative values are further searched with part filters for the likelihood maximization of the presence of occluded object. Commonly occurring occlusion patterns for vehicles parked in streets are utilized by [38] to train part based models for car detection. Authors followed the implementation
and framework of [11] and reasoned that the occluder should also be considered as a part of detection of an occluded object. They modeled occlude-occludee pair for their reasoning and detection in their system. They derived the results for two different models learnt by the images. One of the model is trained such that it learns the parts of occluder and occludee cars as one single object based on commonly occurring patterns in streets. Another model explicitly learns the visible parts of occluded car along with fully visible ones and combines them as mixture of models.

Authors of [23] used pixel gradients around the required object to encode their shape. Gradients calculated on the boundary are assigned to quantized angle values to avoid infinite angle space. Also the norms of these gradients are not considered hence it creates gradient orientation-only maps. They ignore the norm to nullify the effects of contrast changes. Each image gradient does incorporate nearby orientations to allow for local deformation. This makes their method more flexible to intra-class variation. Additionally, they proposed surface normal along with boundary gradients for the applications where depth stream is also available making 3D detection possible.

2.3 Methods for detection of Texture-less Objects

In the past, some studies [40] [21] have been done regarding the modeling of occlusion of objects [26]. One of the ideas introduced was that the regions which are inconsistent in the hypothesized region of object is considered as occlusion [20]. Incorporating these concepts, Hsiao and Hebert [26] proposed modeling of 3D interactions of texture-less objects present in kitchen. In this modeling, visible height and width of the occluded object is reasoned to identify occlusion. They modeled occluder as 3D block with a probabilistic distribution of its presence in front of the required object. Finally they combined object detection methods of LINE2D and Gradient Network with the likelihood of occlusion with prior scene understanding for detection of oc-
cluded objects.

Wang et al. [50] have utilized location and depth information into their hypothesis space. The role of environment is incorporated in the prediction of the presence of objects. This is done in such a way that patches from other objects and background vote for the object being searched for. This, in turn, is a context aware model for occlusion detection. Depth stream obtained by KINECT sensor is used to model layer structure of the scene. Objects present these layers send their vote for the occluded object’s center point and help in segmentation of its area at the same time. Finally a context aware code book is learned in training phase which incorporates contextual information which is necessary for object detection in cluttered environment.

In the case of simpler objects, shape outline based matching techniques are used for simpler objects such as Mugs, cups, swans, bottles etc. For shape outlines, edges are calculated which provide a boundary between the object and its background. One piece of significant work is presented by Ferrari et al. [14]. They introduced the concept of $k$ adjacent segments (kAS) which is a descriptor of local shape property in an image. Object’s boundary is obtained by calculating the edges in a given image and selecting approximately straight pixels in form of segments. Local shape is captured in form of relation between these segments such as angle, size and distance between them. These properties are saved in form of shape descriptor. Segments are considered adjacent when an approximately straight segment ends toward another nearby segment. Depending upon application any number (k) of segments can be considered but an increased number of segments are not suitable as they become more over fitted leaving generality behind. Their descriptor is translation invariant since it does not rely on an absolute position of the segment. Also, it is scale invariant when normalized with respect to the average distance and size of segments. Authors have shown that there are finite numbers of arrangements between segments that can be learned with in a dataset and SVM classifier can be trained on kAS.
In [15] Ferrari et al. took into account a pair of adjacent segments (PAS) and introduced another method for shape learning. Instead of training a classifier by SVM, shape is learnt by a non-rigid point matching algorithm and Hough voting space for counting the PAS found in training image. They claimed to have better results as compared to HoGs introduced by Dalal and Triggs [7]. In [54] PAS is used to extract foreground features which are then combined to obtain root and part models. The root model covers the entire shape while part models cover fine detail of object. Spatial relation is learned between root and part models. These models are combined with sparse coding in form of descriptor. An average of all descriptors is calculated and stored as a final model with which test image descriptors are matched and classified as true or false based on some threshold.

Shotton et al. [43] used contour segments from the edge map of an object to learn a codebook which incorporates fragment detail along with spatial arrangement. Their procedure incorporates the benefits of Bag-of-Words but, unlike of it, they also learn spatial arrangement of the contour segment as a structure of the object. They use chamfer matching which allows the segments to be deformable hence allowing intra-class variations to be incorporated into model. Similarly authors of [29] have also used the similar concept of contour segments but they use AND-OR graphs to accumulate votes for the presence of an object in a detection window. Their framework is arranged into layers in which the bottom layer consists of leaf nodes which have contour segment details. Nodes have inter connectivity to validate consecutive segments and to remove any odd-one-out segments accidentally selected by a node or due to noise. Multiples of these leaf nodes are connected to the layer of OR nodes above them. Multiples of these nodes allow for some segments to be missed due to noise or other background effects. These OR nodes collect the presence of any contour segments to be passed to the layer of AND nodes which combines overall detected segments into a descriptor. SVM classifier scores the detection as true or false. They also incorpo-
rated the concept of mixture-of-models by training the system with multiple views of the same object.
Chapter 3

Methodology

In this chapter, the implementation and framework of deformable part models with HoG features are reviewed and discussed [7] [12] [10]. The benefit of deformable part models (DPM) is that they allow for shape deformation which is necessary to incorporate intra-class variation. Most of the work done for detection of texture-less objects has utilized shape edges as features to train their models. In this thesis occlusion of textureless objects is dealt with DPM and Histogram of oriented gradients (HoG) are selected as features to train the deformable models of texture-less occluded objects. The downside of selecting edges as features is that they present a binary map on the edges of the objects while discarding other texture information present on them. Occlusion further hides the actual edges of the object that makes the detection more difficult by hiding the only-available edge-information. In addition to hiding the actual edges, the false edges of the occluding object do appear in front of the actual object. In this situation, the benefit of HoGs is that they capture rich information about image patch by recording texture information in the form of gradient orientations. That is why the HoG features seem more suitable to detect texture-less occluded objects.
3.1 Concept of Image Convolution, Correlation and inner product

In the coming sections, a very common but an important mathematical operation of correlation is used. This concept is explained here with respect to correlation in 2D images. In image processing, filter, kernel or mask is an n-dimensional rectangular matrix that is convolved with an image or feature matrix to apply some effects such as correlation, blurring, sharpening detecting edges etc. The size and dimension of kernel depend on the applications. In this thesis convolution with gaussian kernel is used to smooth the images. Some of the 2D filter example are given below.

\[
\begin{pmatrix}
1 & 2 & 1 \\
2 & 4 & 2 \\
1 & 2 & 1
\end{pmatrix}
\quad \frac{1}{16}
\begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{pmatrix}
\quad \frac{1}{9}
\begin{pmatrix}
-1 & 0 & 1 \\
-1 & 0 & 1 \\
-1 & 0 & 1
\end{pmatrix}
\]

Correlation, in image processing, is usually used to get similarity measure between images. Some times the image is transformed and to measure its resemblance with the original one, correlation is used. In this thesis image pixels are not directly used rather correlation between features obtained from pixels is used to score the presence of an object at some location. Correlation between two matrices \( f \) and \( h \) at \((i,j)\) is shown below:

\[
f \otimes h = \sum_{k} \sum_{l} h(i + k, j + l)f(k, l)
\]  

(3.1)

Correlation is basically derived from sum of squared differences (SSD). The lesser the value of SSD the more is the similarity between two matrices with zero being same matrices. Let 'f' and 'h' represent two image patches on which similarity is being calculated as given below:
\[ SSD_{\text{min}} = \sum_k \sum_l (f(k, l) - h(i + k, j + l))^2 \]  
\[ (3.2) \]

\[ SSD_{\text{min}} = \sum_k \sum_l (f(k, l)^2 - 2h(i + k, j + l)f(k, l) + h(i + k, j + l)^2) \]  
\[ (3.3) \]

Since the first and the last terms do not provide any information regarding the similarity between patches, rather, these will give squared sum of individual pixel values of both patches. However the middle term is meaningful because it gives the product of the entries belonging to two matrices. so, for similarity measure, middle term should be kept while discarding the other terms.

\[ SSD_{\text{min}} = \sum_k \sum_l -2h(i + k, j + l)f(k, l) \]  
\[ (3.4) \]

By changing negative sign with the positive, SSD subscript is changed from minimum to maximum.

\[ SSD_{\text{max}} = \sum_k \sum_l 2h(i + k, j + l)f(k, l) \]  
\[ (3.5) \]

\[ SSD_{\text{max}} = \sum_k \sum_l h(i + k, j + l)f(k, l) \]  
\[ (3.6) \]

Finally the constant is also removed giving rise to a similarity based on entries of two matrices. This is known as Correlation. For the applications where sizes of filter and features window are same, we can get rid of multiple summations and can convert the multiplication as dot product of two vectors \( f', h' \). This is done by concatenating rows of matrix into a single row and dot product give the result of correlation. This correlation is used as dot product between model and feature pyramid later in this chapter.
3.2 Histogram of Oriented Gradients

Histogram of oriented gradients (HoGs) were introduced by Dalal and Triggs [7]. The computation of HoG features involve three steps (i) gradient calculation at each pixel (ii) aggregation of pixel gradients-features into cells of square shape (iii) normalization of blocks, each consisting of four cells. In first step, gradient at each pixel of image is calculated by simple centered difference mask [-1 0 1] in x and y direction which are then combined to get 2D gradient. As the angle can take infinite continuous values so the angle values should be quantized to some finite number of quantized values \( Q \). Authors recommended 9 quantized values from 0-180 degrees of angle for unsigned gradients or 18 quantized values from 0-360 degrees for signed gradients. Equation 3.7 and 3.8 give the mathematical expression to get the right bin \( B \) for a given gradient angle \( \theta \) at pixel \((x, y)\) with total number of quantized values given as \('b'\) [7] [11].

For signed gradients:

\[
Q(x, y) = \text{round off} \left( \frac{b \cdot \theta(x, y)}{2\pi} \right) \mod b \quad (3.7)
\]

For unsigned gradients:

\[
Q(x, y) = \text{round off} \left( \frac{b \cdot \theta(x, y)}{\pi} \right) \mod b \quad (3.8)
\]

we can call these angle-quantized gradients as pixel-level-features. In the second step, these pixel level features are aggregated into square cells of height \( k \) that give cell features \( C(i, j) \) such that \( i \) and \( j \) vary as \( 0 \leq i \leq \lfloor (w-1)/k \rfloor \) and \( 0 \leq j \leq \lfloor (h-1)/k \rfloor \) for an image of size \( w \times h \). Each gradient feature at pixel \((x, y)\) can be simply assigned to a cell index \( ([x/k], [y/k]) \).

After the aggregation of pixel features into cells, bins are used to collect the magnitude
Figure 3.1: Representation of cells and blocks and 50% sliding blocks. Aggregation bin are shown on the top right corner
of gradients for each quantized angle value for all pixels in each cell. Number of bins are same as the number of quantized angle values for gradients. In the third step, normalization of aggregation of four cells as blocks is done. A major benefit of HoG is that the effect of additive brightness is canceled by taking the difference in the calculation of the gradient however local contrast affects the gradient magnitude. To cancel the effect of contrast, normalization is done over a broader region by combining 4 cells together rather than that of an individual cell. This aggregation of four cells is called a block. Block traversal is done as 50% overlapping manners to do normalization smoothly. Figure 3.1 shows this concept pictorially. $L_2$ normalization $v/sqrt(v^1_2 + v^2_2 + \ldots vn^{2})$ is suggested by the authors to work better than $L_1$ norm and other variants of $L_1$ (eg. $L_1$-sqrt). $L_2$ normalization factor is given as:

$$N(i, j) = (\|C(i, j)\|^2 + \|C(i + 1, j)\|^2 + \|C(i, j + 1)\|^2 + \|C(i + 1, j + 1)\|^2)^{\frac{1}{2}} \quad (3.9)$$

These four cells belong to a block at location $(i, j)$. After the normalized, these blocks are concatenated to generate the feature vector for the overall image. Dalal and Triggs presented an HoG feature of 36 dimensions which was further reduced by Felzenszwalb et al. into 31 dimensions by principal component analysis(PCA). The modified HoG features of Felzenszwalb et al. are used in this thesis work.

### 3.3 Deformable Part Models

In this section, structure of deformable part models (DPMs) is explained. DPMs incorporate the concepts of feature pyramid, root and part filters. All these concepts and DPM structure are explained below.
3.3.1 Pyramid Representation

The size of objects is not known in single camera images so the concept of scale space is used. In scale space, image smoothing and sub-sampling is done for some finite levels as shown in Figure 3.2. An octave is defined as collection of smoothed and sub-sampled images before reaching down at the image of twice the resolution, in the pyramid, as that of current size. A parameter $\lambda$ defines this number. In this thesis $\lambda$ of 5 is used during training and 10 during testing phase.

In DPMs a pyramid of HoG features $'H'$ is also created from the image pyramid because we will deal with feature space primarily. Let $'F'$ be a filter window of width $'w'$ and height $'h'$ and $p(x, y, l)$ be a position at point $(x, y)$ at level $'l'$ of pyramid. If $\phi(H, p, w, h)$ is the feature vector converted to row vector and filter $'F'$ also converted to row vector then the response of this filter at this location is given as the dot product of filter and feature window:

$$F'.\phi(H, p, w, h) \quad (3.10)$$
3.3.2 Defining Models

Deformable part models are star shaped models which contain a root filter that covers the overall shape of the object and part filters that cover different parts of the object based on the high energy portion in the feature domain (HoGs). Part filters incorporate their relative position with respect to root filter location named as an anchor point. Part filters are applied at twice the resolution in feature pyramid as that of root filter as it gives better result of detection. This process is shown in Figure 3.3. Overall model for an object having \( n \) parts has \( (n + 2) \) entries in the model \((F_0, p_1, p_2, \ldots, p_n, b)\). \( F_0 \) is the root filter, \( p_i \) represents part filter and \( b \) is bias term. Part filter is defined as tuple \((F_i, v_i, d_i)\) where \( F_i \) is the part filter, \( v_i \) is the location of anchor point and \( d_i \) is deformation cost. Deformation cost is four dimensional vector that varies as the part deviates from its ideal position so as to penalize false location of a part.

Instead of using sliding window approach, the response of each filter of the model is obtained on the feature pyramid ‘H’. After obtaining the response, an object hypothesis \( z = (p_0, p_1, p_2, \ldots, p_n) \) is formed by grouping the root and part filter responses from a complete model. Part filter responses are those which are an octave level ‘\( l'\) down in the pyramid as explained before. The score of the model at a particular location in the feature pyramid is sum of the responses of root and part filters and the deformation cost is subtracted to incorporate deformation but penalizing response from wrong locations.

\[
\text{score}(p_0, p_1, \ldots, p_n) = \sum_{i=0}^{n} F_i \phi(H, p_i) - \sum_{i=1}^{n} d_i \phi_d(dx_i, dy_i) \quad (3.11)
\]

where
Figure 3.3: Image and Feature pyramids are shown side by side. It can be seen that part filters are applied at twice the resolution as that of root filter

\[(dx_i, dy_i) = (x_i, y_i) - (2(x_0, y_0) + v_i)\]

\[\phi_d(dx, dy) = (dx, dy, dx^2, dy^2)\]

The score function can be expressed in form of dot product \(\beta.\psi(H, z)\) instead of summation as given below:

\[\beta = (F'_0, F'_1, ..., F'_n, d_1, d_2, ..., d_n - 1, d_n, b)\]

\[\psi(H, z) = (\phi(H, p_0), \phi(H, p_1), ..., \phi(H, p_n), -\phi_d(dx_1, dy_1), -\phi_d(dx_2, dy_2), ..., -\phi_d(dx_n, dy_n))\]
3.3.3 Matching

The process for finding the best location of the root filter, the maximum score from parts placements is required. The score is given as

\[
Score(p_0) = \max_{p_1, p_2, \ldots, p_n} score(p_0, p_1, \ldots, p_n)
\]  

(3.16)

In this case the root window is obtained by the best placement of the part responses as opposed to a sliding window approach where parts are obtained with respect to root filter window. The process of finding the best placement is done by finding the response of a specific part filter in the neighborhood of the expected location. Let \(R(i, l)\) be the response of the \(i^{th}\) filter in the \(l^{th}\) level of the feature pyramid then the maximum score of this filter in the neighborhood of location \((x, y)\) is given as:

\[
D_{i,l}(x, y) = \max_{dx, dy} (R_{i,l}(x + dx, y + dy) - d_i \phi_d(dx, dy))
\]  

(3.17)

This response incorporates the deformation cost. Now the overall response of the model hypothesis by incorporating the best part locations is given as:

\[
score(x_0, y_0, l_0) = R_{0,l_0}(x_0, y_0) + \sum_{i=1}^{n} D_{i,l_0 - \lambda}(2(x_0, y_0) + v_i) + b
\]  

(3.18)

\(\lambda\) represents the number of levels that should be moved down in the feature pyramid to get response of part filters at twice resolution. This process is shown in figure 3.3

3.4 Machine learning

Support vector machine (SVM) based classifier is trained on HoG features for model training. Support vector machine (SVM) has shown significant performance in classification tasks. Classification performance depends on the location of decision boundary
between classes. SVM uses a mathematical formulation that choses a boundary which maximizes the margin between classes. This maximum margin boundary improves the classification results for SVM in contrast with the other classification techniques such as perceptron and bayesian classification. These methods provide approximately good solution without a guarantee of global minimum for error value. However SVM provides global minimum when its objective function is minimized by optimization techniques. Objective function of linear SVM is a convex problem but detection of part based objects with minimum information about the location of the object has latent information. So Latent SVM (LSVM), introduced by felzenszwalb et al. [11] [12] is used in this thesis work which has semi convex objective function. The objective function becomes full convex after the latent values are defined for the training examples. Now an overview of the formulation of LSVM is presented.

The objective function of a linear SVM is given below:

\[ L_D(\beta) = \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{n} \max(0, 1 - y_i f_\beta(x_i)) \]  

(3.19)

Here \( D \) represents the training data \( (< x_1, y_1 >, ..., < x_n, y_n >) \) and the last term is Hinge Loss. The score function \( f_\beta(x) \) depends on implementation of SVM. Thus, the score function of a classical SVM is given as

\[ f_\beta(x) = \beta.\phi(x) \]  

(3.20)

which is linear in \( \beta \) making the hinge loss, as well as, over all objective function of equation 3.19 convex. While the score of LSVM is given as

\[ f_\beta(x) = \max_{z \in Z(x)} \beta.\phi(x, z) \]  

(3.21)
where $z$ represents the latent values of the training examples from the set of possible latent values $Z(x)$. This scoring function is maximum of objective functions that are linear in $\beta$. The Hinge loss given in objective function of equation 3.19 is convex for negative examples. But it is, usually semi convex for the positive examples because of the latent values which are not known at the beginning. This objective function becomes fully convex when the latent values are specified for the positive training images.

3.5 Optimization of Objective Function

As mentioned before, latent values for the positive training images should be specified to make the objective function of the LSVM convex. Let $Z_p$ represents the set of latent values for positive images dataset $D$. An auxiliary objective function can be defined that will help in optimization of original objective function of equation 3.19 as

$$L_D(\beta, Z_p) = L_{D(Z_p)}(\beta)$$

(3.22)

Here the data is redefined with specified latent values, $Z_p$, for positive examples. Now the original objective function can be related to auxiliary function as,

$$L_D(\beta) = \min_{Z_p} L_D(\beta, Z_p)$$

(3.23)

This represents that the objective function of LSVM is equal to the auxiliary function minimized over the latent values specified for positive training images. As this is a function of two variables $\beta$ and $Z_p$, coordinate descent method can be used to minimize auxiliary objective function. Following are the two steps of coordinate descent:

1. The auxiliary object function is optimized by selecting $Z_p$ such that it gives maximum score for each positive example.
2. The auxiliary objective function is optimized over the weight matrix $\beta$ when latent values for data are defined. Because objective function will be convex only after the definition of latent values.

### 3.5.1 Defining Latent Values

Latent values depend on specific application for which LSVM is used. In the case of object detection in images, latent values could be defined as the location of the object $L_r$, its bounding box coordinates $S_r$, location of its parts $L_p$, number and size of parts $S_p$, anchor point of parts $A_p$ etc. The set of latent values is given below:

$$Z_p(x_i) = \{S_r(x_i), L_r(x_i), N_p(x_i), S_p(x_i), L_p(x_i), A_p(x_i), H(x_i)\}$$  \hspace{1cm} (3.24)

### 3.5.2 ComputationS of $\beta$

To obtain $\beta$ in second step of the optimization process, at least two methods can be used, namely 1) Quadratic programming 2) gradient descent. There are many tools available to solve quadratic programming problems including Matlab. In this work, stochastic gradient descent is used to calculate $\beta$. At the very first stage, only gradient descent formulation is shown. By taking the sub gradient of the objective function of equation 3.19, we obtain the slope of the function given as

$$\nabla L_D(\beta) = \beta + C \sum_{i=1}^{n} h(\beta, x_i, y_i)$$  \hspace{1cm} (3.25)

Sub gradient is calculated because of the Hinge loss, as this makes the function non differentiable. So intervals are to be defined for this sub gradient, given as.

$$h(\beta, x_i, y_i) = \begin{cases} 0 & \text{if } y_i f_\beta(x_i) \geq 1 \\ -y_i \phi(x_i, z_i(\beta)) & \text{otherwise} \end{cases}$$
Calculation of $\beta$ by equation 3.25 takes a lot of time because it requires a sum over all training examples. To deal with this complexity, stochastic gradient descent is applied by approximating the summation to a multiplication as given below:

$$\sum_{i=1}^{n} h(\beta, x_i, y_i) \sim nh(\beta, x_i, y_i)$$  \hspace{1cm} (3.26)

After all this detailed explanation, the overall algorithm is stated below that is updating the $\beta$ in the second step of the optimization process.

1. Let $i$ be an image in the training dataset
2. Let a learning rate be selected as $\alpha_k$ for the $k_{th}$ iteration of the optimization
3. Let $z_i$ be those latent values that maximize the score function given as:

$$z_i = \arg\max_{z \in Z(x_i)} \beta \cdot \phi(x_i, z)$$  \hspace{1cm} (3.27)

4. Now $\beta = \beta - \alpha_t \beta$ if $y_i f_\beta(x_i) \geq 1$ as stated in the intervals of sub gradients above.
5. Otherwise $\beta = \beta - \alpha_t (\beta - C ny_i \phi(x_i, z_i))$

Here the learning rate is chosen as $\alpha = \frac{1}{k}$ as authors in [41] show that it works well for SVM training.

### 3.6 Data mining of training data

The training of support vector machine classifier involves positive as well as negative examples from dataset. With respect to training examples of images, positive images define the location of the required object however, the case of negative images is different. We can extract many negative examples from a single negative image by
considering different locations with in an image. To avoid training by large number of negative examples, it is a common practice to use the hard negatives as training examples of negative instances. Hard negatives are those negative images that are classified as true by the classifier. So if a classifier is trained on hard negatives, it should also work on easy negatives.

Working on the similar lines as that of hard negatives, LSVM uses the concept of hard examples to obtain convergence by less number of training examples. Hard examples are those instances from training data that are mis-classified by the classifier during training phase. This training process uses similar approach as that of boot strapping in which random sampling and replacement in training data is done to measure the accuracy. First the data mining for classical SVM is described, after that, the description of LSVM is stated.

### 3.6.1 Data mining in Classical SVM

For the training case of classical SVM with input $x$ from training data $D$ and output label $y$, hard and easy examples are defined mathematically as:

$$H(\beta, D) = \{ <x, y> \in D \mid yf_\beta \leq 1 \}$$

(3.28)

$$E(\beta, D) = \{ <x, y> \in D \mid yf_\beta \geq 1 \}$$

(3.29)

It is desired to get $\beta$ with the a subset $C$ of original dataset $D$ rather than full dataset such that,

$$\beta(C) = \beta(D)$$

(3.30)

In the learning process, some examples are loaded and training is done on those examples. Then easy and hard examples are found in those initial examples. Now easy examples are removed and hard ones are added and training cycle is repeated. So
in this process training and example updating are done alternatively. The algorithm is given below:

Let $C_1$ be initial intake of examples from the dataset $D$

1. $\beta$ is learned on some subset $C_k$ for some $k_{th}$ iteration
2. If all hard negatives of the dataset $D$ are in the subset $C_k$ then break the loop
3. Otherwise exclude the easy examples from the current subset
4. Add more examples in the remaining subset of images with at least one new hard example.

Training subset is shrunk by removing easy examples and expanded by adding more examples from dataset with at least one hard example. When all hard examples are added in the current subset then the process terminates.

### 3.6.2 Data mining in Latent SVM

For latent SVM, latent values of positive examples $z \in Z(x)$ should also be included along with the training examples. These latent values are first defined in the training process then the training is done otherwise the solution will not converge. So the optimization is done for the objective function $L_{D(Z_p)}(\beta)$ where $D(Z_p)$ represents the training data with latent values specified. In the case of machine training by features $\phi(x, z)$ extracted from images, these features are included instead of original images. In short, training data is converted into features and arranged such that the pair $(i, v)$ represents the image number in the range $1 \leq i \leq n$ and feature vector $v = \phi(x_i, z_i)$. This converted set of training images into features can be represented by $F$ that contains the pairs $(i, v)$ as explained before. Based on this $F$, the optimization
Based on this redefinition of objective function, gradient descent is also redefined which is adapted to new notation as,

1. Let a learning rate be selected at $\alpha_k$ for the $k$th iteration of the optimization
2. Let $i$ be an instance in $F$
3. $v_i = \arg\max_{v \in V(i)} \beta.v$
4. Now $\beta = \beta - \alpha_t \beta$ if $y_i(\beta.v_i) \geq 1$
5. Otherwise $\beta = \beta - \alpha_t (\beta - Cny_i v_i)$

The objective of training is same as explained before that we want the efficient training with less number of training examples. So if $\beta = \arg\min_\beta L_F(\beta)$ then the $\beta$ obtained after the training process should be $\beta(F) = \beta(D(Z_p))$. Now following the similar manners hard and easy examples are defined as,

$$H(\beta, D) = \{(i, v) \in F \mid z_i = \arg\max_{z \in Z(x)} \beta.v \& y_i(\beta.v_i) \leq 1\}$$

$$E(\beta, D) = \{(i, v) \in F \mid y_i(\beta.v_i) \leq 1\}$$

One major difference in the training data for classical SVM and LSVM is that the latent values set $Z_p$ for positive examples need to be specified. This is represented as $D(Z_p)$. The training algorithm is given below:

Let $F_1$ be initial intake of features obtained from the dataset $D(Z_p)$

1. $\beta$ is learned on some subset $F_k$ for some $k$th iteration
2. If all hard negatives of feature dataset $F$ are in the subset $F_k$ then break the loop.

3. Otherwise exclude the easy examples from the current subset.

4. Add more examples in the remaining subset from $F$ with at least one new hard example.

### 3.7 Model Training

In this section, model structure, preparation of training data and extraction of latent values is explained.

#### 3.7.1 Preparation of Training data

Training data is available in the form of images. For positive images, bounding box coordinates are given in annotation file for each image. There are 1000 negative images of random scenes for training of LSVM. From the bounding box area, the area of root filter is decided. A mean of all bounding box area is calculated and the size of root filter is selected such that it is not greater than 80% of that value. This avoids the over estimation of the root filter size. Now the part filter selection is explained. Number of part filters are pre-defined (i.e. six) and based on this number, root filter area is selected into rectangular portions. This selection is made to cover high energy regions of the root filter in HoG domain. This energy is calculated simply by L2-norm of HoG features. Once an area is selected, its energy is considered zero so as to avoid re-selection of the same area in the next part selection. As explained earlier in this chapter, part are selected at twice the resolution in feature pyramid as that of root filter.
3.7.2 Defining latent values

The location of root and part filters, their areas, their relative position, anchor points etc. are the latent values. When the location of the object is defined along with root and part filters on feature pyramid $H(x)$, this gives redefined training data with configuration parameters of the model. The training algorithm is briefly outlined below:

Dataset Images: Positive Images $= \{(I_1, B_1), (I_2, B_2), \ldots, (I_n, B_n)\}$ Negative Images $= \{N_1, N_2, \ldots, N_m\}$

1. Convert positive images to feature domain

2. Train root filter based on bounding box with random negatives.

3. Get part filter regions and related latent values. Then repeat training with hard negatives obtained from step 1.

4. Do data mining by removing easy examples and adding hard negatives

5. Repeat the loop for finite iterations due to practical reasons such as time efficiency.
Chapter 4

Experimental Results

In this chapter, results of this research work are presented. A new dataset of 1000 images of mugs and cups is introduced. Along with this, a synthetic dataset of 600 images with occluded mugs having varying degree of occlusion is also presented and made available online\(^1\).

4.1 Dataset

As per the best knowledge of the author, there is neither any publically available dataset of occluded mugs, nor many publicly available datasets of non-occluded mug images. One such dataset is ETHZ Shape classes that contains images from five different classes namely Mugs, Swans, Bottles, apple logo, giraffe, etc.. There are only 48 images of mugs in this dataset. There is no one correct answer of how many training examples should be there to train to achieve the convergence of a support vector machine based classifier. However, training examples should not be too few. To get the appropriate number of images for machine training, well known PASCAL VOC competition \([9]\) was searched for the training images for different categories.

\(^1\)https://drive.google.com/drive/folders/0B3CGFrN2xvVka0Z1Q0pIRUk3dDg?usp=sharing
of objects. Based on the number of the training images of the competition, 1000 images are collected, out of which, 800 are used for training and validation purpose. Remaining 200 images are used to create testsets of varying extent of occlusion as explained in the following items.

- **Images Source:** Images have been collected from Flickr which is a photo collection and sharing website. Flickr has been a source of images for well-known publicly available dataset such as PASCAL VOC challenge [9]. Medium size images are collected with sizes around 640 x 425. A collection of 1000 images contain nearly all poses of mugs except a view from exact top. In order to introduce variation and to avoid biasness towards a particular shape, images of different shapes are present in this dataset. In addition, intra-class variation is also introduced by capturing the mugs of different sizes (height and width). Moreover, there also exists variation in the shapes of cup holder. The rigid shape models do not give optimal solution when such a large intra-class variation is present in the training set. Considering large intra-class variation, the DPMs provide optimal solution as they look for deformable parts of the objects.

- **Synthetic Dataset:** When real dataset is unavailable, researchers, generally, create and evaluate the performance of their algorithm on synthetic dataset. For example, authors of [47] tested their algorithm for occluded pedestrian detection with synthetically created dataset. They cropped people from images and pasted them on other pedestrians to simulate a scene of crowd where people occlude other people. In [36], authors also use the synthetic dataset for human pose estimation. Similarly, authors in [32] [27] have used synthetic datasets for research purposes. Following the similar approach, we have examined the performance of proposed methodology on synthetic dataset as there does not exist a real dataset for occluded mugs and cups. A collection of 1000 images of mugs is used for this purpose. Out of 1000 images, 800 are used in training of
DPMs and remaining 200 are used for creation of test sets. From the 200 images, three datasets are created each with 200 images. The difference between each dataset is the extent of occlusion. These datasets have 25%, 50% and 75% occlusion images of mugs and cups for testing. For the training of models with 800 images, there was no need for creation of three training datasets, instead, only different bounding box coordinates were required for 25%, 50% and 75% occlusions. That is why, in this case, different annotations are generated having bounding box with above mentioned occlusion levels.

Commonly found items are used to occlude the objects in this dataset. To avoid sharp changes at the borders of the occluding objects, alpha blur is used to smooth the transition and give a natural look to the images as shown in Figure 4.1. This figure shows some examples from all three synthetic datasets of occluded mugs. These datasets are available online.

- Ground truth Bounding box: For all of the images in the dataset, annotation XML files are also provided that contain information about image ID (for give credit to photographer), dimension, ground truth bounding box and color channels.
4.2 Test procedure

4.2.1 Precision-Recall

In order to evaluate the performance of the algorithm, precision-recall curves are used in this thesis. This thesis is related to object detection in images so this discussion is done with this contextual information. True positives are those instances that actually are objects correctly detected by the system while false positives are those instances which are not the required objects but are detected as true by the algorithm. Similarly true negatives are those detected objects that are labeled false by the system and these are actually false. False negatives are system misses of true positives. Precision is the ratio of detected true positives over the sum of true and false positives. Recall is the ratio of all detected true positives over all true positives. Figure 4.2 shows graphical representation of this concept.
4.2.2 Criterion for true detection

The criterion of true detection considered in this thesis is intersection over union (IoU) of 50%. In the IoU an intersection of proposed bounding box and actual ground truth bounding box is calculated as shown in Figure 4.3.

Some authors choose lenient criterion of 20% such as [15] but somewhat strict criterion of 50% is used for testing in this thesis. The reason for following this strict criterion is because this is one of the major challenges to draw bounding box as accurately on the object as possible. Now the results are presented for the testing of proposed method on testsets.
4.3 Test results

In order to extensively evaluate the performance of proposed algorithm, three sets of experiments were conducted: 1) with prior knowledge of extent of occlusion and 2) without prior knowledge of extent of occlusion 3) rule based detection of extent of occlusion. Testing is performed to get the optimal results of detection for occluded objects. Now the methodology to conduct these experiments is briefly explained. First of all, each of the three models is applied on the test set of the same degree of occlusion as of the training set for models i.e. model trained on 50% occlusion is applied on the testset of 50% mug occlusion images. This is done to get the maximum value of average precision having a prior knowledge of degree of occlusion. Next, each model is applied to a mixture of images without any prior knowledge about the image under test. In the third experiment, all models are applied on each image and a rule based reasoning is used to get the boundary box according to the extent of occlusion. Details are given below:

4.4 Testing with prior knowledge of occlusion

In this section we discuss the results of models when degree of occlusion is already known. A model trained on a particular occlusion is applied on the testset of the mugs with the same occlusion level. We obtain three PR curves for the three models. This gives us the maximum average precision (AP) of detection of a model when occlusion level is known. Precision Recall curves for detection of 75%, 50% and 25% occluded mugs is shown in Figure 4.4. It can be seen that Average Precision is around 70%. It is observed that AP is not very high because of the very simple shape of the mug which is further occluded by other objects. However it can still be considered reasonable.
Reader can observe that AP of all three cases is and close together. One of the interesting information can be observed at lower end of the curves. Recall value are almost same (Recall \( \sim 0.80 \)), at those point, for all three cases but the precision values decrease slightly for 75% occlusion as compared to 25% occlusion. If we look at mathematical representation, precision is given as \( \frac{TruePositives}{TruePositives + FalsePositives} \) and recall as \( \frac{TruePositives}{TotalPositives} \). Constant recall value means that the detected true positives are same in all cases so the decreasing precision value means that false positives start appearing making the ratio go down. That seems quite intuitive too because increasing level of occlusion adds more noise and leaves less information, about the object, which invites more false detections. Besides this small decrease in precision at lower end of curve, average precision is same for 25% and 75% occlusion.

The AP of 50% occlusion case is higher than the other two cases because this case is most suitable for the detection of occluded objects as explained in the next section of detection results with out any prior occlusion knowledge. Results are summarized in Table 4.1.

### 4.5 Testing without prior knowledge of occlusion

In this section we discuss the results of model when there is no prior knowledge about degree of occlusion of the object. It is observed in the previous section that applying a particular occlusion model to the images with known degree of occlusion gives fairly high value of average precision. This is not, however, a practical approach because in real life, the degree of occlusion is not known. To make the scenario more practical, it is tried to find the optimal solution to occlusion detection of texture-less objects with no prior knowledge. An experiment is done to see the behavior of models when each of the models is applied on a mixture of all 600 test images with no prior knowledge about any particular image. The reason to select this approach is that a trained
Figure 4.4: PR curve for DPM based occluded texture-less object detection with prior knowledge of degree of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results.
model on a particular extent of occlusion can detect objects with different degree of occlusion. That is because it is the visibility of the object which is changing and not the object itself. So to observe the behavior of each model, this approach is tested. It is observed that the average precision of 25% and 75% model has decreased to 39.2% and 44.1% respectively. But an interesting and useful result here, to observe, is that the AP of 50% model is 63.7% which is fairly high on the mixture of data with no prior knowledge about occlusion. PR curves are shown in Figure 4.5.

The deterioration of performance for 75% models is due to the IoU criterion. Ideally, this model should detect almost all occlusion case as it is trained on least visible mug images. But, actually, due to strict IoU criterion of 50%, this model failed to register detection on 50% and 25% occlusion images. While the low performance of 25% occlusion model is due to two reasons. First is due to false negative detections as this model is trained on most visible portion of mug images. So this rejects the mugs with 75% occlusion which results in false negatives. Second is, again, due to IoU criterion.

The performance of 50% occlusion model is still very high as compared to other models. This is because this detects all three degrees of occluded objects and the bounding box suggested by the model gives better IoU with ground-truth than the other two models. The AP by this model is 63.7% which is close to the average of three APs (72.8%) of results obtained with prior occlusion knowledge. Results are summarized in Table 4.1

### 4.6 Testing with detection heuristics

It is known to the reader that detection results of models with known occlusion are considerably high. It is tried to combine the models in such a way that they could provide comparable results on unknown occlusion with those results which are ob-
Figure 4.5: PR curve for DPM based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results.
tained with occlusion knowledge. To achieve this, an approach is experimented in which all three models are applied on each image and the decision about the degree of occlusion is obtained based on score of these models. It is explained earlier that more than one model can vote a detection because all three models are trained on mugs with the only difference of extent of occlusion. Based on highest score of the model, the bounding box is assigned to a detection because the detection results also depend on better selection of bounding box. Rules for the selection of bounding box based on the score of the detection are given below. It is observed, however, that even with the rule based reasoning for the estimation of occlusion, the maximum AP of 59.5% is achieved which is lower than that of 50% occlusion model applied to mixture of images. The PR curve is shown as 4.6.

![PR curve for DPM based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results](image)

Figure 4.6: PR curve for DPM based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion test results
Rules: Models are named M25, M50 and M75 with respect to 25%, 50% and 75% occlusion models. Calculate difference between score obtained and minimum threshold which is named as $\Delta T_{75} \Delta T_{50} \Delta T_{25}$

If (only single vote is received from any single model and rest votes are zero)
Then (assign that model bounding box)
Else if (only M75 and M50 voted)
Then if ($\Delta T_{75}$ is greater than $\Delta T_{50}$ by 1)
Assign BB of M75
Else assign M50
Else if (only M05 and M25 voted)
Then assign M50
Else if (all models voted)
if (($\Delta T_{75} - \Delta T_{50}$) $\leq$ 0.3 and ($\Delta T_{75} - \Delta T_{25}$) $>$ 1)
Assign M50
if (($\Delta T_{75} - \Delta T_{50}$) $>$ 0.3 and ($\Delta T_{75} - \Delta T_{25}$) $>$ 1.5)
assign M75 if (($\Delta T_{75} - \Delta T_{50}$) $\leq$ 0.3 and ($\Delta T_{75} - \Delta T_{25}$) $<$ 1)
assign M25
if (($\Delta T_{75} - \Delta T_{50}$) $\geq$ 0.3 and ($\Delta T_{75} - \Delta T_{25}$) $<$ 1)
assign M25
else assign M50

Table 4.1: Average Precision of DPM with HoG features

<table>
<thead>
<tr>
<th>Degree of occlusion</th>
<th>With prior occlusion knowledge</th>
<th>Without prior occlusion knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% occlusion</td>
<td>71.2%</td>
<td>39.2%</td>
</tr>
<tr>
<td>50% occlusion</td>
<td>75.9%</td>
<td>63.7%</td>
</tr>
<tr>
<td>75% occlusion</td>
<td>71.3%</td>
<td>44.1%</td>
</tr>
</tbody>
</table>
4.7 Comparison with edge based method

Comparison of results is the results of detection with edge based method, the well known work of Ferrari et al. is selected because their implementation is available to public. First of all, a brief overview of their methodology is presented.

- Their methodology of object detection is based on edges of the objects. They introduced the concept of \( k \) adjacent segments (kAS) which is a descriptor of local shape property in an image [14]. Object’s boundary is obtained by calculating the edges in a given image and selecting approximately straight pixels in form of segments. Local shape is captured in form of relation between these segments such as angle, size and distance between them. These properties are saved in form of shape descriptor. Segments are considered adjacent when an approximately straight segment ends toward another nearby segment. Depending upon application, any number of segments can be considered but increased number of segments are not suitable as they become more over fitted leaving generality behind. Their descriptor is translation invariant since it does not rely on absolute position of segments. Also it is scale invariant when normalized with respect to average distance and size of segments. Authors have shown that there are finite number of arrangements between segments that can be learned with in a dataset.

They further presented their method with pair of adjacent segments PAS (\( k=2 \) in kAS) proposing that two segments perform well for shape learning in [15]. In this paper they introduced a shape learning with a non-rigid point matching algorithm and Hough voting space for counting the PAS found in training image. They claimed to have better results as compared to HoGs introduced by Dalal and Triggs [7].

Figure x shows two same images but the detection results of different approaches
are shown. It is observed that the edge based approach missed the clearly visible object because the edges of the occluding object successfully created false edge to deceive the detector. While the DPM did not miss the object because it is not relying on shape outline rather the overall shape gesture.

For comparison of the results from these two methods, it was necessary that the training and testing datasets should be same. So first of all, pair of adjacent segments (PAS) were generated for all datasets and models were trained for all 25%, 50% and 75% occlusion datasets. Testing results are presented in the Figure 4.7 for detection with prior knowledge of occlusion and Figure 4.8 for detection without prior knowledge. The IoU criterion of 50% is also kept same in this case. It can be observed that the AP in this detection methods is considerably low as that of DPMs. Results for edge based method are summarized in Table 4.2

Table 4.2: Average Precision of edge based object detection method

<table>
<thead>
<tr>
<th>Degree of occlusion</th>
<th>With prior occlusion knowledge</th>
<th>Without prior occlusion knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% occlusion</td>
<td>33.1%</td>
<td>28.1%</td>
</tr>
<tr>
<td>50% occlusion</td>
<td>27.9%</td>
<td>30.7%</td>
</tr>
<tr>
<td>75% occlusion</td>
<td>35.8%</td>
<td>33.3%</td>
</tr>
</tbody>
</table>
Figure 4.7: PR curve for edge based occluded texture-less object detection with prior knowledge of occlusion. Curves are in order of 25% (Top), 50% (middle) and 75% (bottom) occlusion testset results.
Figure 4.8: PR curve for edge based occluded texture-less object detection without prior knowledge of occlusion. Curves are in order of 25%(Top), 50%(middle) and 75%(bottom) occlusion occlusion testset results
The reason for the low precision of the edge based method is due to more false positives detected in images. And these false positives is due to the presence of cluttered edges of other objects. To show the effectiveness of the proposed methodology, Figure 4.9 is presented which shows the results of detections from the two methods. False detections of object can be seen on the left side by edge based algorithm while the correct detections are given on the right side by DPM trained on HoG features.

Figure 4.9: Images on the left show detection by deformable part models with red bounding box as overall detection window. Images on the right show false detection by edge based method on same images.
Chapter 5

Conclusion and Future Work

In this chapter, conclusions of this research work are presented. Along with that, future directions for this type of research, in particular, the 3D detection of texture-less occluded objects by deformable part models is considered.

5.1 Conclusion

In this thesis the problem of detection of texture-less occluded objects is examined. Occlusion of objects happens almost everywhere around us; so it is an important area of research. To deal with the occlusion problem the methodology used in this thesis is a combination of HoG features, deformable part models and latent support vector machine (LSVM). The reason of adopting this approach is because this is more suitable to detect texture-less objects. The detection results obtained through this approach are compared with a published edge-based approach for detection of texture-less objects. Target object chosen for this research work is coffee mug/cup. Varying degree of occlusion is taken into consideration namely 25%, 50% and 75% however the handle is visible in all cases. There is no publicly available dataset of occluded mugs/cups so the images of mugs/cups are collected from flickr and the occlusion of object is created synthetically. This dataset is made available online to be used for the research work
in this area.

To evaluate the performance of the algorithm on the detection of texture-less occluded objects, precision-recall (PR) curves are used. A PR curve is a well acknowledged representation to evaluate the performance of the detection algorithm. The PR performance of the proposed algorithm is compared with the performance of edge based algorithm and our results are clearly better than the other methodology. Following results are extracted from the series of experiments performed on occluded mugs:

- Experiments with prior knowledge of occlusion show that the HoG based DPMs outperformed the edge based method in all three levels of occlusion namely 25%, 50% and 75%. Higher precision means that out of all detected instances, there were more true positives than false positives. While the less precision in case of edge based method means that the mug instances were detected at the places where there was no actual presence of the object. This implies that the occlusion introduces more edges that could imitate the presence of the object at false locations so deceiving edge based algorithms. While the improved results for our methodology tackled that situation in a better way.

- As the prior knowledge of occlusion is not the practical case so the experiments were performed on a mixture of test images with all extents of occlusion. Each DPM model was applied separately to the combined testset and it is concluded that the model trained on 50% occlusion performed best from the models trained on 25% and 75% occlusion examples. Although it is 12% than that of with prior knowledge but it is still very high as compared to edge based models.

- All the models are trained on the mugs but with different extent of occlusion, so in the next set of experiment, it was tried to judge the extent of occlusion in an image by applying all models to a single image. Rules were defined to assign a bounding box of that occlusion extent whose model’s score is the maximum
out of all three. However the maximum score obtained in this case was still 4% less than that of the 50% model.

Based on the results from the experiments it is concluded that HoG based DPMs outperform edge based method and are more suitable in the detection of texture-less occluded objects. Finally, the generalized solution is the model trained on 50% occlusion for the detection of occluded mugs because it performs best on the unknown extent of occlusion. It is therefore concluded that this approach provides better detection of texture-less occluded objects and requires further investigation.

The limitations of the work include the texture-less mug object chosen as occluded object. It needs to be explored if the proposed methodology works for all kinds of texture-less objects including jugs, plates, pots etc. Other than that the size of dataset also matters for the training of the SVM. Currently, number of images in the training dataset are based on the research work done by other researchers in other SVM training research on images [9].

5.2 Future Work

With improving technology, there are devices that are appearing in the market (Kinect sensor) which provide depth information along with 2D RGB image making it possible to get 3D information of the scene. 3D detection of texture-less objects by deformable part models could be more useful in the detection because depth information could be added in decision making process. With depth information, shape of an object can be registered as per its actual presence.

Machine trained classifiers will also incorporate depth information during training phase and the detection score will also depend on the obtained depth information. With depth information, occluding object can easily be detected in front of occluded object which is not possible in case of 2D image. Deformable part based detection will
further strengthen the detection score which could be made close to human perception. Other texture-less objects can be tested with the proposed methodology to see the effectiveness of detection algorithm.
Bibliography


