OULUN YLIOPISTO
UNIVERSITY of OULU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Irtiza Hasan

BENCHMARK EVALUATION OF OBJECT SEGMENTATION PROPOSALS

Master's Thesis
Degree Programme in Computer Science and Engineering
July 2015

Hasan I. (2015) Benchmark Evaluation of Object Segmentation Proposals. University of Oulu, Department of computer Science and Engineering. Master's Thesis, 57 p

ABSTRACT

In this research, we provide an in depth analysis and evaluation of four recent segmentation proposals algorithms on PASCAL VOC benchmark. The principal goal of this study is to investigate these object detection proposal methods in an un-biased evaluation framework.

Despite having a widespread application, the strengths and weaknesses of different segmentation proposal methods with respect to each other are mostly not completely clear in the previous works. This thesis provides additional insights to the segmentation proposal methods. In order to evaluate the quality of proposals we plot the recall as a function of average number of regions per image. PASCAL VOC 2012 Object categories, where the methodologies show high performance and instances where these algorithms suffer low recall is also discussed in this work. Experimental evaluation reveals that, despite being different in the operational nature, generally all segmentation proposal methods share similar strengths and weaknesses. The analysis also show how one could select a proposal generation method based on object attributes.

Finally we show that, improvement in recall can be obtained by merging the proposals of different algorithms together. Experimental evaluation shows that this merging approach outperforms individual algorithms both in terms of precision and recall.

Key words: object detection, object recognition, object segmentation proposals

ABSTRACT

Table of Contents

1. INTRODUCTION 6
   1. 1 Introduction 6
   1. 2 Scope of Thesis 7
   1. 3 Structure of Thesis 7
2. OBJECT DETECTION 9
   2.1 Overview 9
      2.1.2 Benchmarks and Evaluation 13
      2.1.3 Background and Current State-of-the-Art 20
   2.2 Candidate Regions for Object Detection 24
      2.2.1 Sliding Windows 24
      2.2.2 Object Segmentation Proposals 25
3. OBJECT DETETCTION PROPOSALS 26
   3.1 Geodesic Object Proposals (GOP) 26
   3.2 Multiscale Combinatorial grouping 28
   3.3 Proposals with Global and Local Search 30
   3.4 Selective Search 32
   3.5 Past Evaluation of Object Segmentations Proposals 33
   3.6 Summary 33
4. EXPERIMENTAL EVALUATION 35
   4.1 Segmentation Proposal Recall 35
   4.2 Evaluation Protocol 35
   4.3 Results at Instance Level 35
   4.4 Results at Class Level 40
   4.5 Results with Combination of Different Algorithms 41
5. DISCUSSION AND CONCLUSION 52
6. REFERENCES 54

PREFACE

This thesis was carried out at Center for Machine Vision University of Oulu, Finland under the supervision of Dr. Juho Kannala and Dr. Esa Rahtu. I would like to express my deepest gratitude to Dr. Kannala for continuously guiding me and his presence has been a source of inspiration for me. I would also sincerely like to thank Dr. Rahtu, his feedback and suggestions have been the critical driving force in my thesis. My regards to Mr. Saad Ullah Akram from CMV for sharing his views and enriching me with the fundamental knowledge of this domain.

Finally, to my parents and brother whom I suppose, I cannot thank enough. Without their early support and continuous guidance, none of this would have been possible.

Oulu, 29th July 2015

Irtiza Hasan

ACRONYMS AND ABBREVIATIONS

BING	Binarized Normed Gradients for Objectness Estimation
CPMC	Constrained Parametric Min-Cuts
CIODC	Category Independent Object Detection Cascade
GOP	Geodesic Object Proposals
MCG	Multi Scale Combinatorial Grouping
PGLSV	Proposals with Global and Local Search
RIGOR	Reusing Inference in Graph Cuts for generating Object Regions
SGDT	Signed Geodesic Distance Transform
SS	Selective Search
TRAINVAL	Training and Validation

1. INTRODUCTION

1. 1 Introduction

Automatic object detection is a fundamental task in computer vision. Many researchers have investigated this topic in detail [1, 2], in the recent decade and the outcome is promising. Traditionally regarded as a classification problem, sliding window methodology is the most popular approach in this paradigm. Classical approaches [3, 4] in object detection are exhaustive methods, classifiers are evaluated in a sliding window manner over all locations and on all possible scales. A well-known disadvantage of this approach is in order to achieve a higher accuracy, number of bounding boxes must be extremely large and a lot of computational time is spent on worthless boxes. Due to enormous number of operations, the applied classifiers are quite time consuming to evaluate especially when the number of object classes is large. Even for a relatively limited dataset such as [5], the computational cost of these traditional methods is not practical. After the rise of the several segmentation and recognition datasets such as Pascal VOC [6], new state-of-art algorithms have focused widely to address this challenging issue. A new approach, object segmentation proposals has placed itself at the front, in object detection paradigm. This new wave of segmentation driven object detection is quite promising.

This alternative approach uses segmentation methods to extract candidate regions which can be used as an input to different image analysis and classification methods. Object segmentation proposal are a recent development in object detection. The development of the field has led many researchers to propose new methods. In a nut shell the idea behind detection proposals is to generate few thousand candidate regions per image with high confidence level that the generated proposals cover most of the objects present in the given image. These proposals are class-agnostic and this approach is computationally efficient. These segmentation driven techniques now allow researchers to develop more robust and efficient classification algorithms. Figure 1a shows the overall system of object detection.

This massive shift in paradigm from sliding windows to segmentation proposals is illustrated by the fact that three [7, 8, 9], of the top ranked algorithms on ILSVRC [10] and Pascal [6] are segmentation driven. Most traditional approaches train a specific detector for each class and it affects the overall performance of the object detection pipeline. The accuracy of segmentation-related proposal algorithms are enhanced by the fact that the need of labelling out every pixel that whether this belongs to an object or not is not required anymore. In contrast it generates multiple object proposals with high probability that they will contain the object of interest. However, in segmentation proposals one of the decisive factor is that all or most of the object should be covered in generated proposals, since the objects that are missed would not be detected at all. A higher recall is essential in object segmentation proposals. It is even speculated [11] that segmentation proposals could increase the accuracy of detection as well since the number of false positives would decrease dramatically in comparison with sliding windows approaches. Considering the future research in object detection it is therefore important to evaluate methods that are

reported to be computationally inexpensive. The principal goal of this thesis is to evaluate different methods which achieve high recall and to investigate possible improvements that could be made to these high performing algorithms, which eventually would increase the precision of object detection pipeline?

1. 2 Scope of Thesis

The main contribution of this groundwork is the benchmark evaluation of different segmentation proposal methods. This research also conduct a deep analysis of how these methods perform when they are compared against each other, and what are the factors when these algorithms do not reach up to the expectation. This thesis provides an unbiased evaluation of four state of the art algorithms and gives an insight on how they perform in a unified evaluation. Additionally, this research also carried out an evaluation when two techniques were merged together. The experimental evaluation revealed that it improves the overall recall of the system. This new finding could indeed be a further step in object detection pipeline and could also help in classification tasks.
In brief the main contribution of this work is as follows

- Reviewing of some of the top performing segmentation proposal techniques and evaluation.
- Drawing a comparison graph for these proposal algorithms based on specific parameters.
- Analyzing algorithms at class level, in order to get a better insight on which categories algorithm perform better or have constraints.
- Based on the analysis, combining proposals of two or more than two algorithms together to achieve a higher recall at instance and class level.

It is important to mention here that for some algorithms such as [12] they have not published their results on class level whereas others [8, 13, and 14] have reported it differently on their papers. There was an urge to see their results at class level in order to get a more detailed understanding of their limitations which was kept in mind while merging different techniques together.

1. 3 Structure of Thesis

The organization of this thesis is as follows.

Chapter 2 provides a background of research that is being conducted in this field and how this field has evolved over the period of time. It explain different methods of both bounding boxes and segmentation proposals. Methods that perform well but are not included in this research are also briefly explained here. The purpose of this section is to give researchers an introduction of the history that how this relatively new field has evolved over time and what are the recent trends. This section also provides other avenues to explore for the future work. Algorithm Description, chapter 3 presents each algorithm. In chapter 4 Experimental Evaluation, the approach used for evaluating methods and the results obtained from this methodology is brought in discussion. Finally chapter 5: Discussion and Conclusion summarizes the interpretation of results

furthermore it is discussed that what meaningful conclusions can be drawn. Lastly some future research activities are enlisted that are planned to be carried out.

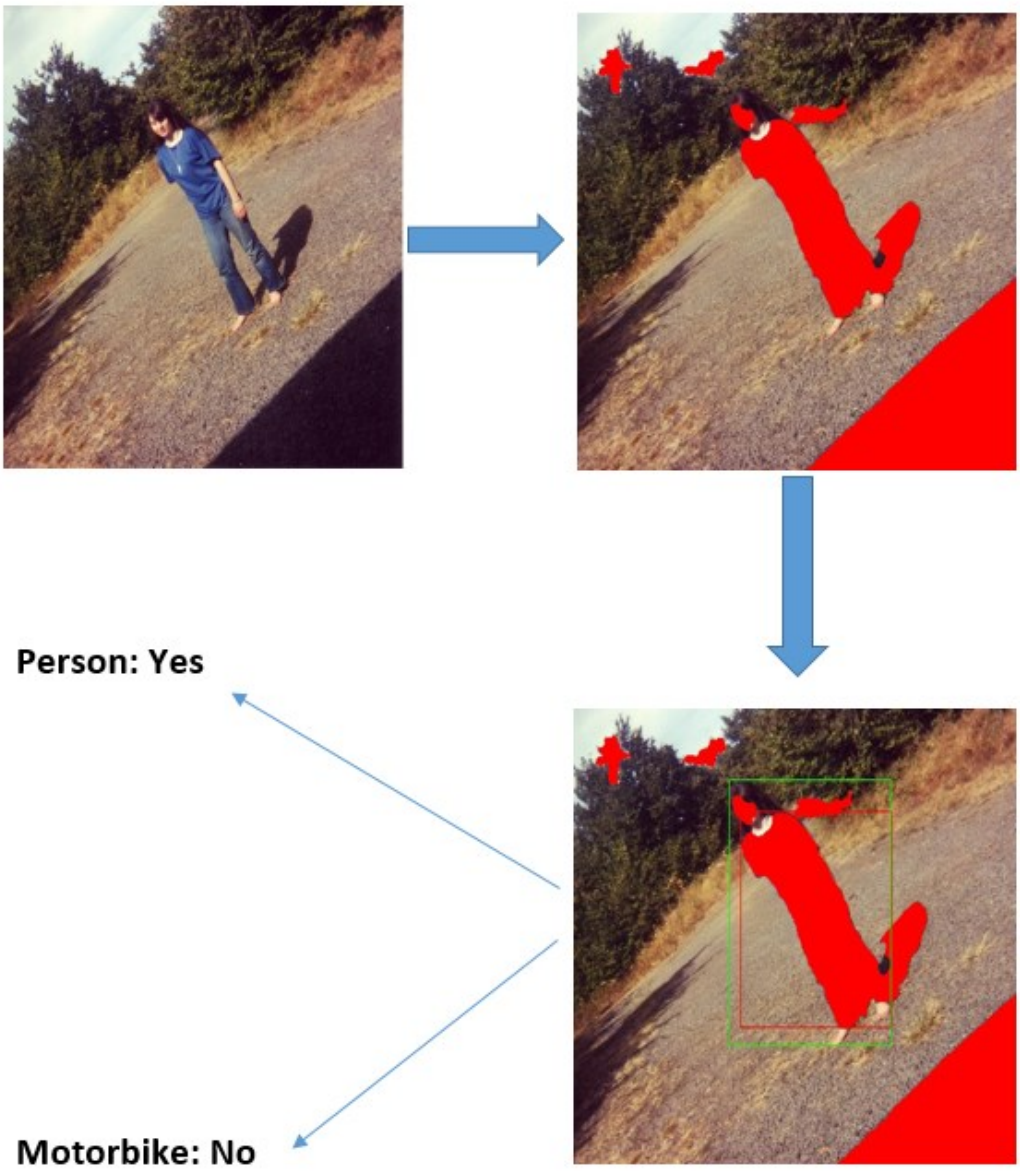

Figure 1a: A general Object Detection system.(1) Top left: Input image is passed to the pipeline.(2)Top Right: Proposals are generated using object segmentation proposal methods.(3) Bottom right: Object detection is performed, Red bounding box depicts object detected by proposal method, Green depicts ground truth.(4)Bottom Right: Finally, classification is performed using machine learning methods.

2. OBJECT DETECTION

2.1 Overview

Object detection and recognition is one of the longest standing problems in computer vision. Acquiring important semantic from images is the foundation of numerous segmentation and classification methods. Briefly, most object detection methods comprise of a feature extraction stage, paired with machine learning architectures which provide meaning to the derived and quantified features. Before going into the operational detail of object detection and recognition few of the problems that are common in this paradigm are shown in Figure 2.

Figure 2: Some common issues in object detection and recognition. See text for detail. Image courtesy of Antonio.

In most real world scenarios the object of interest is often hidden or require some form of preprocessing that in some context enhances its detail in order to be detected by an automatic system. Few of the common issues are depicted in Figure 2. The view point from which the image is taken can vary greatly across images or sequence of frames

and it could lead to a different physical representation of the same object in 2-D plane. Illumination, is also a challenge which could give different meaning to the different parts of the same object or could lead to issues where the part of the object is merged with the background as can be seen in Figure 2, top right the part of the skin due to low illumination got merged with the background. Middle row of Figure 2 shows two problems, middle left the deformable objects i-e non-rigid objects also become problematic, since it is likely that objects would not appear in a constant physical shape. The variation in shape becomes hard to recognize in real life cases. Figure 2 Middle right, in the images besides region of interest there could other details that could or could not be of interest for the recognition system. Being able to extract only the object of interest from a background full of other semantics is arguably the hardest problem in object detection. Somewhat similar to background clutter, problem with occlusion is that some part or even most of the object could be hidden behind which hides or changes the general appearance of the object, which could even be challenging for the naked eye to correctly categorize the object left alone an automatic system as seen in Figure 2 bottom left. Lastly, scale is a problem that is encountered when one deals with the objects that occur naturally in various scales. One of the popular approaches [4] that performed substantially well in early 2000s suffered low accuracy when encountered multi scale objects.

Over the years in object detection and recognition various methods have been deployed to solve the aforementioned problems. However, over the last decade the overall structure of the object detection pipeline has remain unchanged. Although, massive improvements and changes have been made to the individual processing steps. In general the processing steps of the object detection system can be shown by the following block diagram.

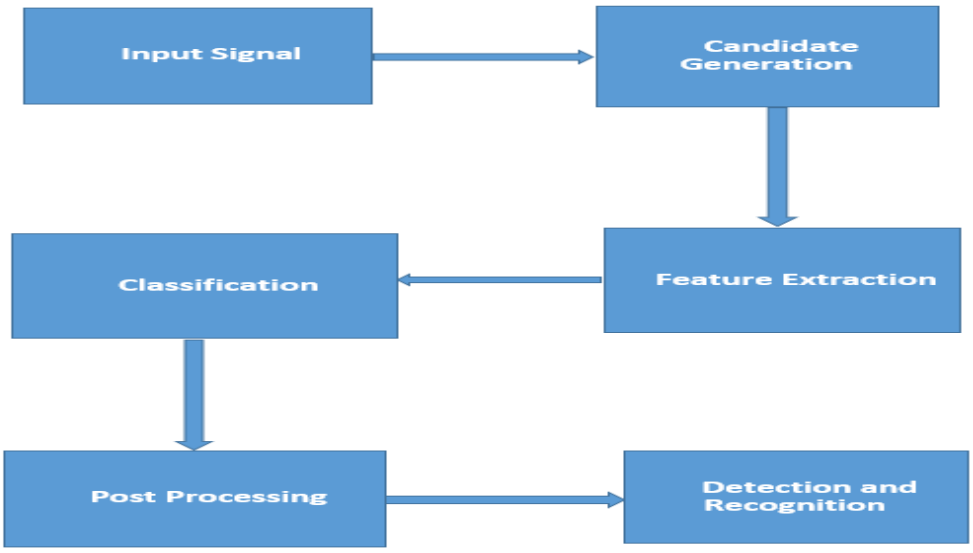

Figure 2b: Block diagram of different processing steps object detection pipeline.

Figure 2b shows various steps in object detection. The **input** to an object detection system is often a 2-d signal or image. It is important to note that often the difficulty of the problem may very well depend upon the limitation of the input signal transmitter. Limitations such as low resolution, distortion and signal to noise ratio etc. These factors could affect the overall accuracy of the object detection system.

Object candidates generation together with segmentation is the backbone of object recognition system. This step determines the precision of the pipeline. The objects that are missed, will not be detected at all by the system. This is the reason, this particular field is a hot topic among researchers. Over the years, various methods have been proposed starting from sliding windows to part based models and eventually to learning architectures. In this module, the algorithms try to find the objects or regions of interest in input image. The working of the algorithms could be very different to each other but in brief, generally based on spatial clues, Color information, Histogram, Saliency and Seed based approaches, the idea is to generate few thousand potential candidates with high probability that they would contain the object of interest.

Ideally, the purpose of **Feature Extractor** is to make job easier for the classifier. The feature extractor tries to characterize an object by measurements whose values are similar for same classes and dissimilar for different classes. It is important that feature extractor utilizes features that are invariant to different transformation, translation and rotation under most cases. As discussed by [15], the image of a simple coffee cup undergoes dramatic variation if it is rotated to an arbitrary angle. In the recent past, a lot of research has undergone in this field and descriptors such as SIFT, SURF and LBP etc. have shown promising results on different benchmarks. However, the process of feature extractor is much more case and domain dependent. A good feature descriptor for one scenario or an object might not be good for another scenario.

Based on the information provided by the feature extractor, the job of the **Classifier** is to assign label or category to each object. Presence of similar features in object of different categories often make the process of classification more challenging. On the other hand the values belonging to a category could be very diverse. This variability could be due to the presence of noise or complexity etc. Hence, a good classifier has the ability to tackle all above mentioned challenges and classify the objects correctly to their respective categories based on available clues.

The purpose of **Post processing** is primarily to perform cleansing on the already classified data. There could be duplicate representation of a same object, in post processing steps often non-maxima suppression is performed to get rid of duplicate detections. Secondly, based on the confidence score in post processing steps some detections are removed as seen in Figure 2c. As depicted after the post processing steps thresholding was performed and all the detection that has below 0.5 confidence were dropped.

Finally, in **Detection and Recognition** the objects are recognized based on the information from previous modules and the output is displayed. It can be seen from the Figure 2c and above discussion that in object detection system, the loss in accuracy in one of the steps effects the overall detection rate.

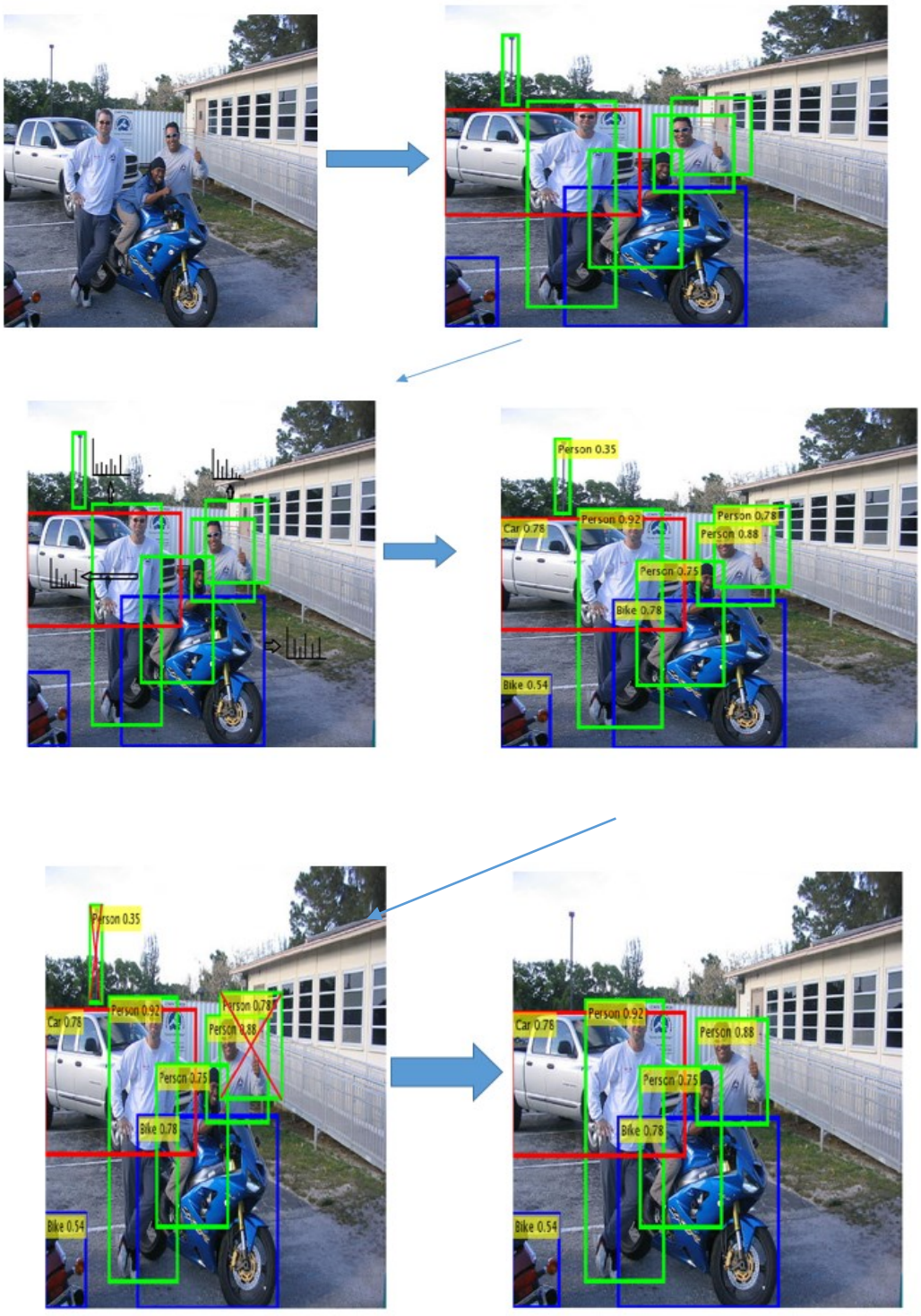

Figure 2c: Practical demonstration of a classical object detection system. Top Left the input image is passed to the system. Top right detection candidates are generated with high accuracy. Middle left features are extracted from each potential object candidate. Middle right based on Features objects are classified and labels are assigned. Bottom Left Non maxima suppression is performed to remove duplicate objects or candidates with low confidence. Bottom right finally the output with correct classes label is displayed.

2.1.2 Benchmarks and Evaluation

All state-of-the-art algorithms in object detection require some form of machine learning. In order to train and test classifiers, datasets with diverse object classes are preferred. Learning architecture such as deep learning rely heavily on large datasets to train accurate classifiers. These issues were partly addressed by datasets such as Caltech 101 [5] and UIUC [16]. However, most of these datasets offered a limited variability in terms of classes and objects. Secondly some of these suffered from the fact that objects occupied maximum portion of the image and objects were present in the center. Thirdly the images were not challenging i-e without occlusion, background clutter and texture. Finally as discussed by [17], there were only one instance of a class per image as shown in Figure 3a. Dataset bias was a huge problem lately in object detection. To solve the aforementioned problems, construction towards diverse and more challenging datasets started. Two of the widely used datasets will be discussed briefly here.

Figure 3a: Top and bottom left are images taken from PASCAL VOC 2012 (Airplanes and Motorbike). While top and bottom right are from Caltech101. Caltech 101 has only one instance of a class per image, whereas PASCAL VOC 2012 has more than one instance of a class per image. PASCAL VOC 2012 has more challenging scenes.

PASCAL VOC [18] is a publicly available benchmark of annotated images. A challenge in visual recognition and funded by PASCAL network of excellence. The competition was ongoing since 2006-2012. The number of object classes is 20, making it a diverse dataset. The quality of the images are better than most of the existing datasets. The images are more complex and have challenging features such as truncation and occlusion. The dataset is divided into two datasets. Training/Validation (Trainval) and Test data (Test). The trainval dataset can be further decomposed into training and validation. However, it is left to the researcher's discretion to use any subset from trainval set.

Each image in the trainval set is carefully annotated with bonding box for each instance of the 20 classes. Additionally, for each object there are some other attributes as well such as "**Difficult**","**Orientation**""**Occluded**""**Truncated**". Figure 3b explain these attributes in PASCAL VOC dataset.

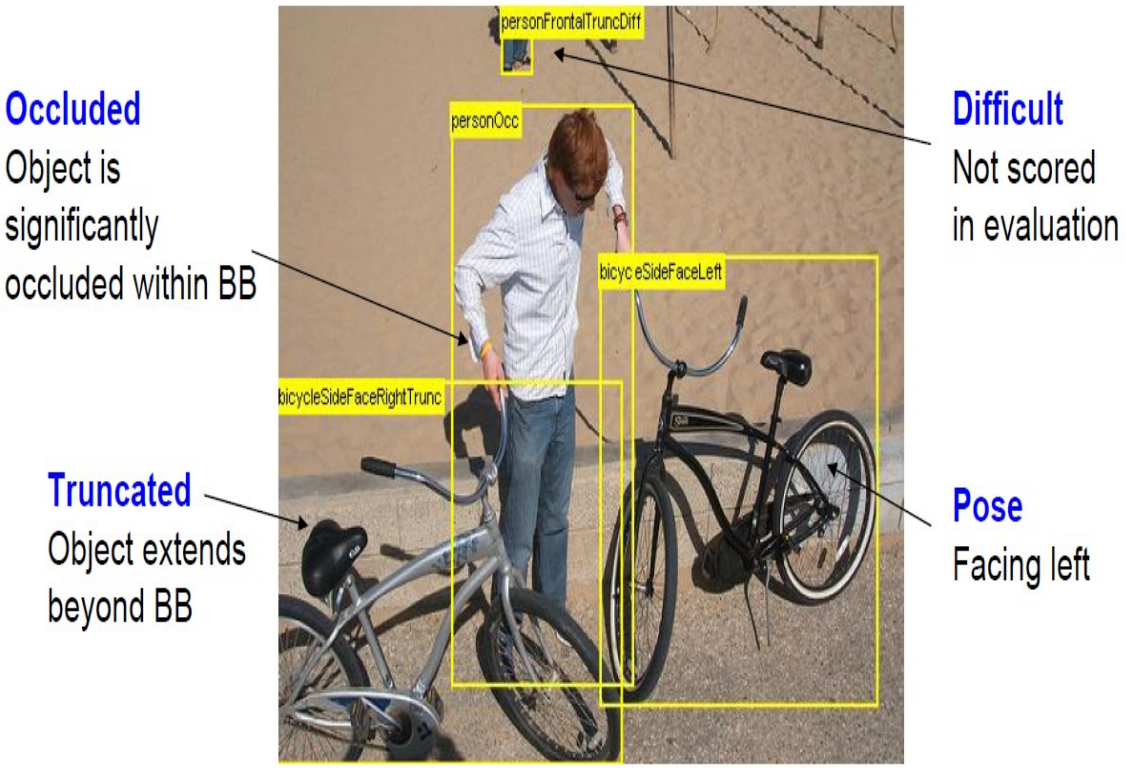

Figure 3b Different Annotation types on PASCAL VOC 2012. Illustrations adapted from slides by Andrew Zisserman.

For PASCAL VOC 2012 primarily there are three challenge tasks. **Classification**, **Detection** and **Segmentation.** Additionally, there are two subsidiary task on **Action**

Classification and **Person Layout.** In a nut shell **Classification** as described by Zisserman et al, is there an ***X*** in the image? **Detection** refers to where are the ***X's*** are in the image? **Segmentation** is defined as which pixel belong to ***X*** ?

Table 1a: PASCAL VOC 2012 classes (20 in total).

Vehicles	**Household**	**Animals**	**Others**
Aeroplane	Bottle	Bird	Person
Bicycle	Chair	Cat	
Boat	Diningn table	Cow	
Bus	Potted Plant	Dog	
Car	Sofa	Horse	
Motorbike	TV/Monitor	Sheep	
Train			

For **Classification,** predict the presence or absence of at least one object of a class form each twenty object classes that are listed in Table 1a, given the input image. Besides prediction, participants are also expected to provide a confidence score. It is the prerogative of the researchers to choose all or any object classes. For example they can choose "Bus" or "Bus and Cars". As explained by [18], two competitions are defined according to the training data. First one is if the training data is chosen from VOC trainval data. Participants are allowed to use the provided annotations for training. However, change in annotations is not permitted. For the second competitions participants are allowed to use any source of Data excluding the VOC test data. Consequently any training data can be used except provided test images. Figure 3c top row illustrates the classification challenge.

Detection refers to the prediction of the bounding box for each object of a class, for each twenty object classes given the input image. Researchers are also required to provide the confidence score. Similarly, participants may chose a particular class or work on all classes of VOC dataset. Like classification, detection also contains two competitions based on the data that is being used. Figure 3 c bottom row depicts detection challenge.

Segmentation is prediction of the object class for each pixel i-e assign a corresponding label to each pixel in the image based on its class or label the pixel as background if it does not belong to any of the twenty object classes. Participants are not required to provide confidence score in this competition. Figure 3d shows the segmentation challenge

Action Classification was introduced in 2010. The idea is for each of the ten action classes listed below, predict if a person enclosed in bounding box is performing that

action or not. A confidence score should also be associated with the prediction. The ten action classes are “jumping”,”phoning”,”playing instrument”,”reading”,”riding horse”,”riding bike”,”running”,”taking photo”,”using computer”. As previously, participants can choose any subset or all of these classes to tackle. Figure 3e explains action classification challenge

Person Layout corresponds to the prediction of presence or absence of parts (heads, hands and feet) along with the bounding box for these parts, Figure 3F. A confidence score is also expected to be provided along with the prediction. Similar to classification and detection challenge there two competitions based on the training data.

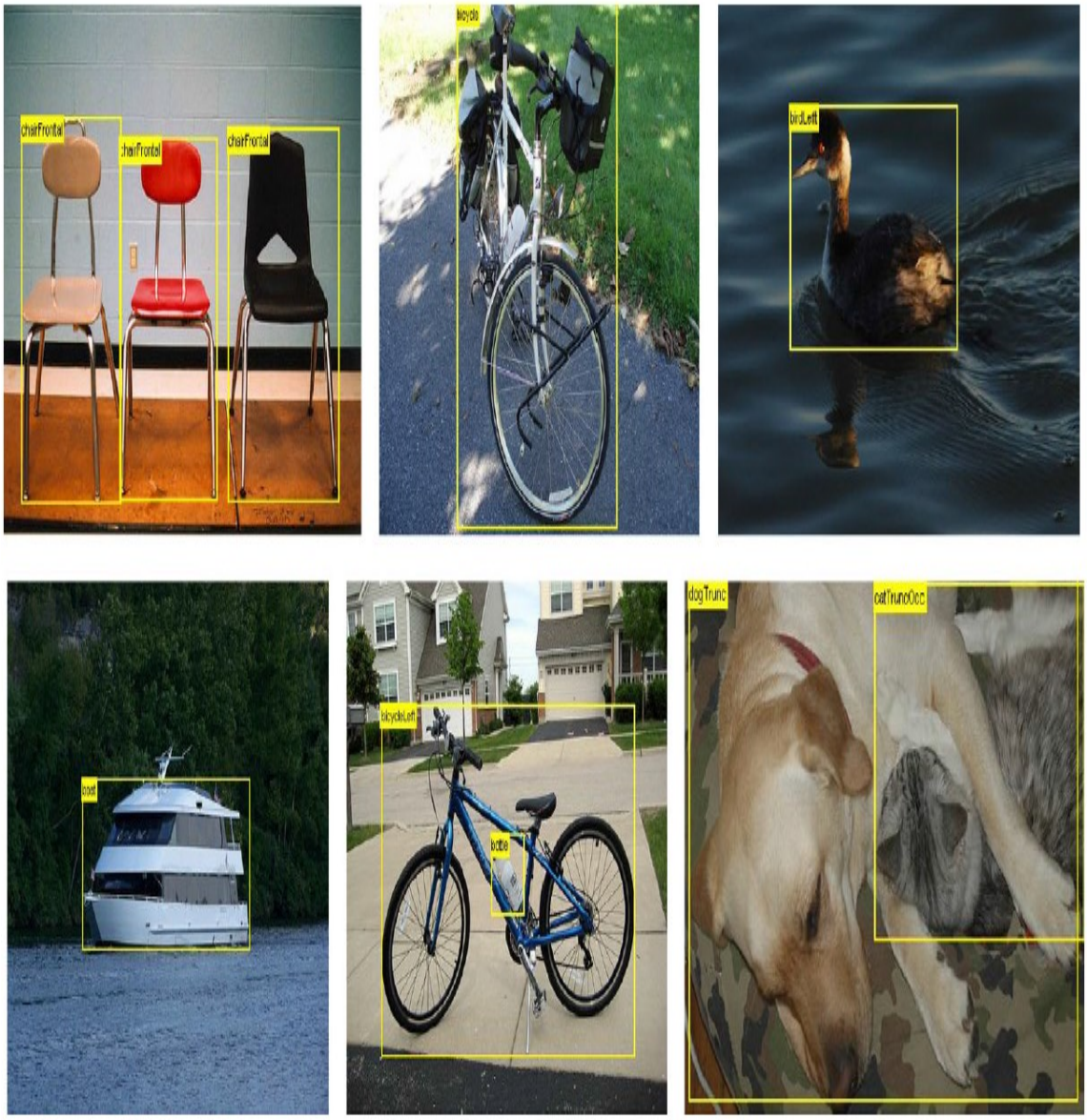

Figure 3c: Top row, classification challenge in VOC 2012 Dataset. Requires prediction of absence or presence of a object from particular class. Bottom row refers to Detection challenge in PASCAL VOC 2012 i-e prediction of the bounding box for each object.

Figure 3d: Segmentation challenge in PASCAL VOC Dataset. Participants are expected to label each pixel(to the class it belongs) or categorize it as background if it does not belong to any VOC class.

Figure 3e: Action Classification challenge. Participants are expected to classify the action that is being performed in the input image.

Figure 3f: Person layout challenge. Participants are expected to predict the presence of head, hands and feet in the input image.

Classification is evaluated by checking whether the predicted class is in given image or not. Detection is evaluated using the jaccard index given in equation 1.

$$J(A,B)=|A \cap B| / |A \cup B| \qquad (1)$$

A refers to the bounding box area of ground truth and **B** refers to the bounding box area of the prediction. Detail working of the equation is provided in chapter 4. Segmentation accuracy is determined by the equation 2.

$$\text{seg.accuracy= true positive/(true positive + false positive + false negative)} \qquad (2)$$

Action classification is evaluated in similar manner as to classification and Person layout is evaluated using equation one for each individual part.

ILSVRC is another benchmark in object detection and recognition. This benchmark is more diverse than [6] both in number of images and object classes. The number of images are millions along with hundreds of object classes [10]. This challenge has been conducted each year since 2010. ILSVRC has a publically available dataset with ground truth annotation, each year a competition is also held and a corresponding workshop. This benchmark further addresses the issue of previous dataset biases in a more comprehensive way. A comparison chart between PACAL VOC 2012 and ILSVRC 2014 is provided in Table 1b.

Similarly to PASCAL VOC this benchmark is also used for evaluation of algorithms in Object detection and Classification. Broadly, the dataset is divided in to two categories.

- **Detection**
- **Classification and Localization**

In the **Detection** dataset there are 200 object categories, in contrast to 20 object categories in PASCAL VOC dataset. However, similar to PASCAL VOC each dataset is further spitted in to three categories Training, Validation and Testing. First two categories are made public for the researchers to develop and test their algorithms whereas the last Testing set is used evaluation of the algorithm by the organizers and this set is not publically available to provide an unbiased evaluation. Each image in the detection set is fully annotated and all the categories are labelled. While collecting the images scale, size, level of image clutterness and number of object instances were considered. These attributes made this dataset as currently the most diverse dataset ion object detection. Figure 3g shows some sample images from detection set of ILSVRC 2014.

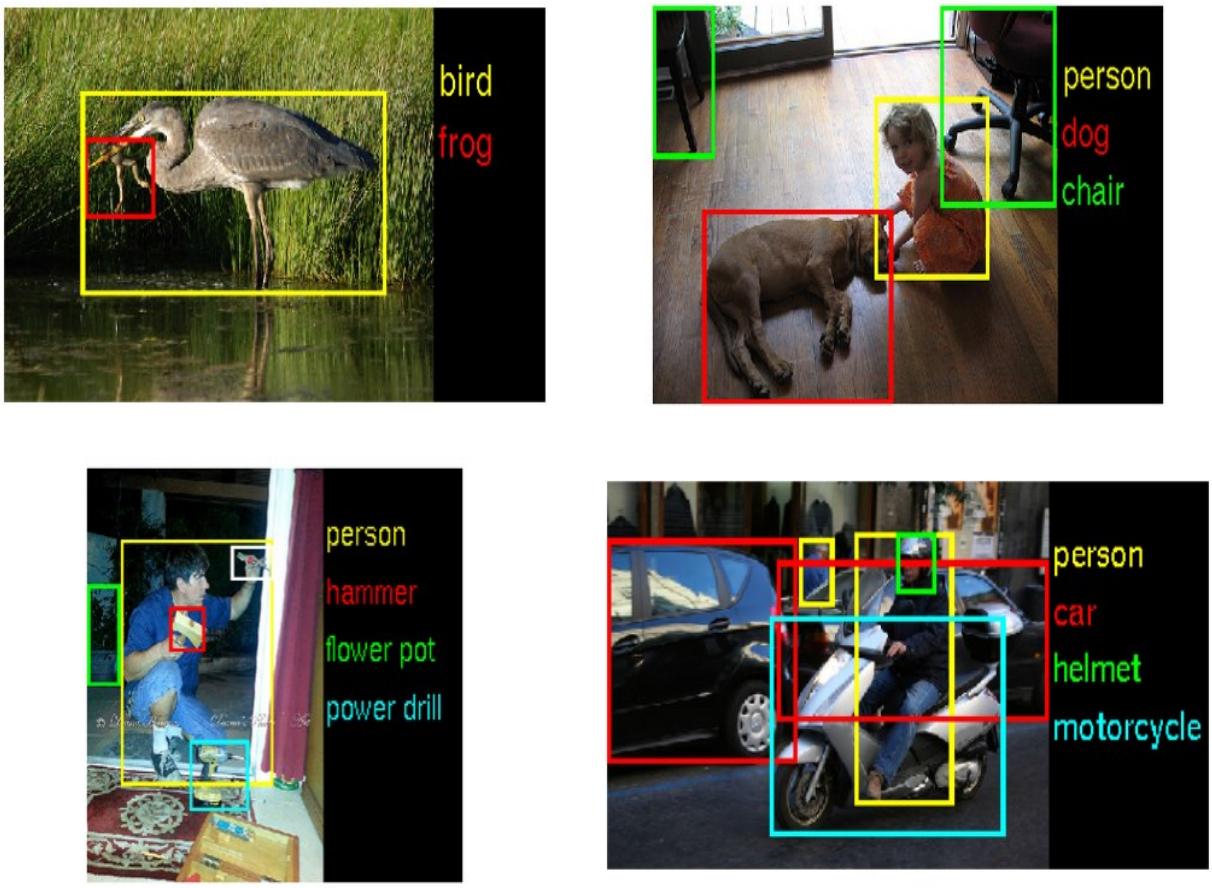

Figure 3g: Some of the example images of ILSVRC2014. Objects could be in several challenging locations such as top left and bottom left. Whereas different object could be very close packed to each other or in some occasion they are overlapping as in top right and bottom right.

There are 150,000 images in Validation and Test dataset of **Classification and Localization** data. These images are then carefully annotated with the presence or absence of 1000 object categories. As of now, 50,000 images with labels are released as

a validation data, list of labels for object categories and a development kit. The rest of images are used for evaluation and they do not contain the labels for object categories.

Table 1b: comparative scale of two datasets. Table taken from ILSVRC homepage.

Number of Object Classes		**PASCAL VOC 2012** **20**	**ILSVRC 2014** **200**
Training	Images	5717	456567
	Objects	13609	478807
Validation	Images	5823	20121
	Objects	13841	55502
Testing	Images	10991	40152
	Objects	-----	-----

Finally, there are two tasks in ILSVRC 2014 dataset. **Detection** and **Classification and Localization.** For the **Detection** task each algorithm is expected to output three set of annotations. Class labels (***Ci***), bounding box that encapsulates the object (***Bi***) and a confidence score (***Si***). For each instance of the 200 object categories the annotation set is expected to contain it. Rationally, objects missed or duplicate object detection are penalized.

Classification and Localization is the second task in ILSVRC 2014. An algorithm is expected to produce 5 class labels in decreasing order of the confidence score and bounding boxes, one for each class. The label that best matches the ground truth image would be considered as the correct label, algorithm will be evaluated based on the best matching label. This would allow researcher to make algorithms capable of identifying multiple objects and algorithms will not be penalized if they identify an object that is in fact present in the image but not included in the ground truth.

2.1.3 Background and Current State-of-the-Art

Earliest reporting of automatic object detection dates back to 1950s and 1960s [19].Initially in late 1950s and early 1960s the concepts from signal processing were widely used in object detection. Concepts such as autocorrelation and template matching were known to have been exploited by the earliest object recognition systems. However, these concepts were soon overtaken in 1970s by 3-D shape representations. Volumetric parts were used for the modelling of objects such as generalized cylinders and superquardics[20].In order to remove the representational gap between models and

images, the community focused to capture images in more controlled environment where illumination, structural detail, scene clutter was tailored for the recognition system. However, as one could imagine the results were not up to the mark when real world conditions were applied.1980s saw the improvement in these representational models. Models inspired from CAD were effective 3-D templates [20].Representational gap was considerably closed down by bringing models close to the object presents in the image. However, still the presence of texture and surface marking seriously affected the model. The computational overhead was quite high as well. This formed the basis of modern recognition methods in computer vision.1990s a major paradigm shift took place where the community tilted towards appearance based recognition instead of model based recognition. The representational gap was vanished when models were brought down to the level of images. Powerful machines also aided in the tilt towards appearance based models. For the first time object with complex structures were also recognized. A time line of object detection can be seen in figure 4a and figure 4b.

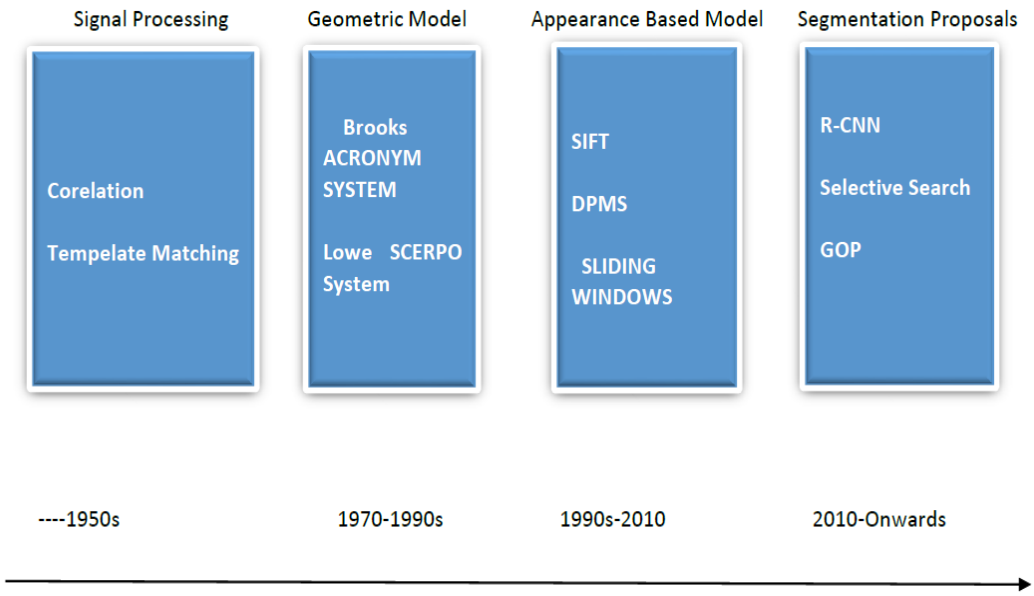

Figure 4a: A block Time Line Diagram of object detection. Initially started from signal processing concepts and evolving to the current state-of-the-art.

Figure 4b: Four decades of object detection. Top left (1970s) Brooks ACRONYM system based on 3-D models (volumetric). Top right LOWE SCERPO system (1980s-mid 90s) used perceptual grouping. Bottom left (late 90s early-2010) scale invariant part based models. Bottom Right (2010- onwards) segmentation proposals and learning architectures. Images adapted from [21, 22, 23, 24, 25, 26].

In the last decade a popular line of research in object detection was to use part-based models in visual recognition tasks. [27], deployed part based model in object detection and showed promising results on state-of-the-art datasets such as [6]. Initially motivated by [28], work called pictorial structures that dates back to 1970s. Basic idea behind deformable part-based models is to represent objects by set of deformable parts. Each part is mapped uniquely, the configuration between the parts is represented by "spring-like" connections. Eventually a classifier is trained for each object. Results on the benchmark dataset showed the improved performance on variable object classes. Before R-CNN revolutionized the object detection paradigm, deformable part-models were considered the top performing methodology on PASCAL dataset. Figure 4c shows the results on one such class from [6].

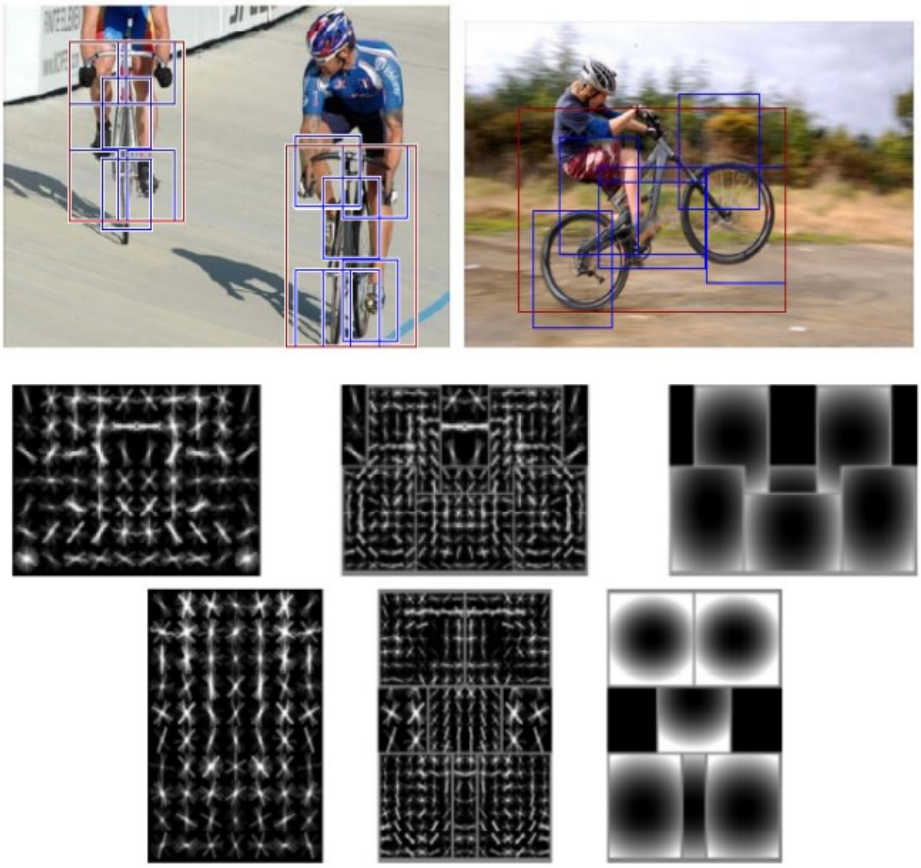

Figure 4c: Results of DPM on **Person** class form Pascal 2007 dataset. Image adapted from [27].

Recently [7], published their scalable method called R-CNN and claimed to have improved the relative accuracy up to 30% on PASCAL VOC 2012. In contrast to the success achieved, the algorithm is relatively simple. It can be categorized in to three parts. Initially class independent region proposals are generated. R-CNN is not constrained by any proposal generation method, in their research they have used Selective Search [8]. In the second step, a high-performance convolutional neural network is applied that extracts 4096-dimensional feature vector from each generated proposal. In their architecture they use five convolutional layers and two fully connected layers. Finally, they train class specific linear SVMs that classify objects. For a limited dataset for training, performance boost was achieved by supervised pre training of the network with abundant data initially and then tailoring the network for actual task with scarce data. Figure 4d shows the overall architecture of R-CNN

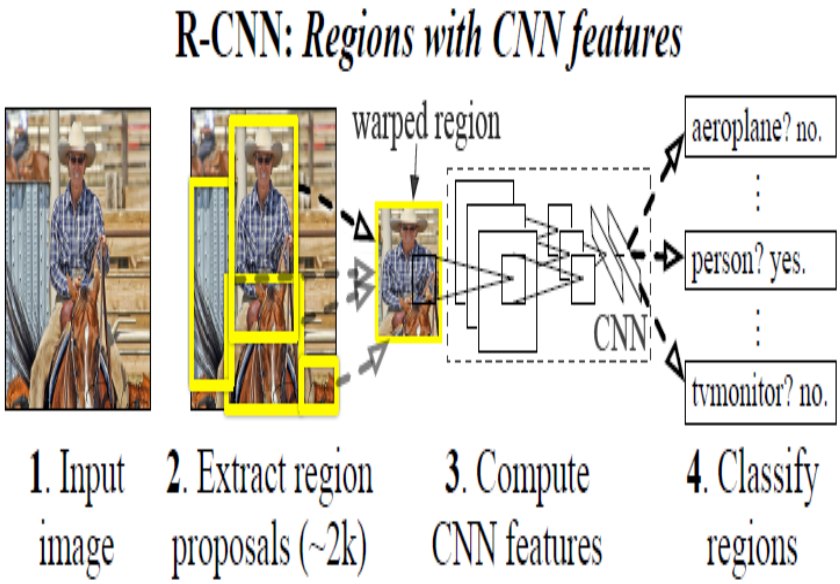

Figure 4d: Showing R-CNN [7] architecture.

2.2 Candidate Regions for Object Detection

Another popular approach to object detection is to use location candidates. The concept behind this paradigm is to generate potential candidates for the object present in the image. Ideally, the number of generated candidates should be low and accuracy should be high. Approaches generally exploits spatial and color clues to generate such candidates. In the recent past, these methodologies have proven to be one of the most successful approaches in object detection.

2.2.1 Sliding Windows

Sliding windows paradigm was once the most popular and successful approach in object detection. [4] The famous Viola-Jones algorithm primarily for face detection, achieved sensational results back in early 2000s and the idea was intuitively simple. Slide a window of a fixed size across the image with a defined step size. Each sliding window was given as an input to the cascade of nodes known as strong classifiers .If a certain condition was met, the window was passed on to the next stage, rejected otherwise. Windows which made it through the complete cascade stage were classified as faces. This idea was fairly robust and ran in real time. Although, Violla-Jones was good at detecting faces, it struggled to detect classes with varying aspect ratios.

HOG, was popularly used for object detection in particular pedestrian detection. Introduced by Dalal et al[3], the idea was given the input image, slide a window across it. The detector window is decomposed into overlapping blocks. HOG feature vectors are extracted in these blocks and then fed to the linear SVM classifier for labelling. This method achieved reasonably good success in pedestrian detection. Though, quicker than Violla-Jones framework HOG speed was still not up to the mark as well

based on the fact that it was a sliding window paradigm. The method was sensitive to occlusion as well.

2.2.2 Object Segmentation Proposals

Object detection paradigm had undergone a considerable shift in last 5 years. The sliding window paradigm was mostly overlooked in the favor of segmentation proposals. Sliding window classifier takes around 10^4 to 10^5 windows per image in a single scale detection as discussed by [29]. Instead of costly sliding windows, object segmentation proposals produce few thousand candidate regions generated with the assumption that they would contain objects present in the image. A brief description of some of the methods is given in [30].

- Objectness[26] is considered one of the earliest work in segmentation proposals. "Objectness" is a term used to define how likely region of interest is an object or not. Based on salient features proposals are generated. These proposals are ranked further according to different cues such as spatial location, superpixels, colors and edges.
- CPMC[25] Generates superpixels by solving graph cut with various random seeds along with unaries applied to pixels in order to obtain foreground and background segmentations. Each generated mask labelled as foreground serves as a proposal. Avoids hierarchical segmentation, Proposals are ranked based on particular features.
- RandomizedPrim[24] Merges low-level superpixels randomly based on a merging functions that compute weights which are learned.
- Chang[31] Saliency, Objectness along with a graphical model is used to merge initially generated superpixels.
- RIGOR[32] Improved version of CPMC, it uses the previous computations across different graph cuts problems and eliminate many redundant computations.
- EdgeBoxes[33] Method computes scores for the windows based on object edges. In order to improve the recall the author propose fine tuning and non-maxima suppression for any desired overlap threshold.
- CIODC[34] Superpixels are used for object proposals along with pairs and triplets. An efficient scoring strategy is proposed which makes improvements to the objectness[26] up to 10 percent improvement in recall rate.
- BING[35] A fast class-independent detector is obtained simply by training a linear classifier over edge features. Then the classifier slides across the whole image similar to sliding window.

There are other methods for object proposal that perform well but are not discussed here. Additionally, the four methods used for evaluation in this thesis are also not discussed here. They are brought to light in upcoming sections.

3. OBJECT DETETCTION PROPOSALS

For this study, we in depth analyze four object detection proposals algorithm. The reason for selecting these four methodologies was the reported high performance on challenging benchmarks. These four algorithms are quite dissimilar to each other in terms of design. Therefore, the analysis of strengths and weaknesses of these four algorithms was performed. Before going into the experimental detail the brief description of the algorithms and past evaluation of object segmentation proposals is discussed in this section

3.1 Geodesic Object Proposals (GOP)

Philip Krähenbühl and Vladlen Koltun[12] proposed technique called Geodesic Object Proposals. The main motivation behind the research was to formulate a method that generates fairly accurate proposals (high recall) with less computational time. GOP produces the least amount of proposals (using the default settings). However, number of proposals can be increased by changing the settings provided in the code.

The approach they presented in their paper represented in Figure 5a can be divided into four stages. Initially given an input image **I**, the aim is to decompose the image into superpixels. Label the edge of each super pixel with a boundary probability map represented as a weighted graph. The boundary probability image is computed using structured forest approach.

The second step is seed placement, main goal is to hit maximum objects with small number of seeds, which would result in less number of proposals. Reducing the computational time for the recognition stage. The proposed learning based seed approach outperforms other heuristics such as saliency based, random or regular seed placement. Seed features exploited by the classifiers are absolute spatial coordinates, normalized coordinates, color covariance between superpixel pixels and seed pixels and finally geodesic distances to previously placed seeds and image boundaries.

In the penultimate phase, foreground and background masks were generated from each seed. Initial approach for mask generation was to label each seed as a foreground and image boundary as background mask. However, this approach was further improved by a learning based approach. Features used by classifiers for mask generation were, location relative to the seed, distance to the image boundary edges and color similarity in multiple color spaces.

Finally signed geodesic distance transform (SGDT) is computed for both background and foreground masks over the image. The geodesic distance between two nodes is defined as the length of the shortest paths between the nodes in geodesic space. It is important to note that although, every level set of SGDT corresponds to a region, but not every region is a good proposal. Good proposals, are extracted by identifying the particular critical level sets (stationary points in geodesic function) of the SGDT. Eventually non-maxima suppression is done to remove any near duplicate object proposals.

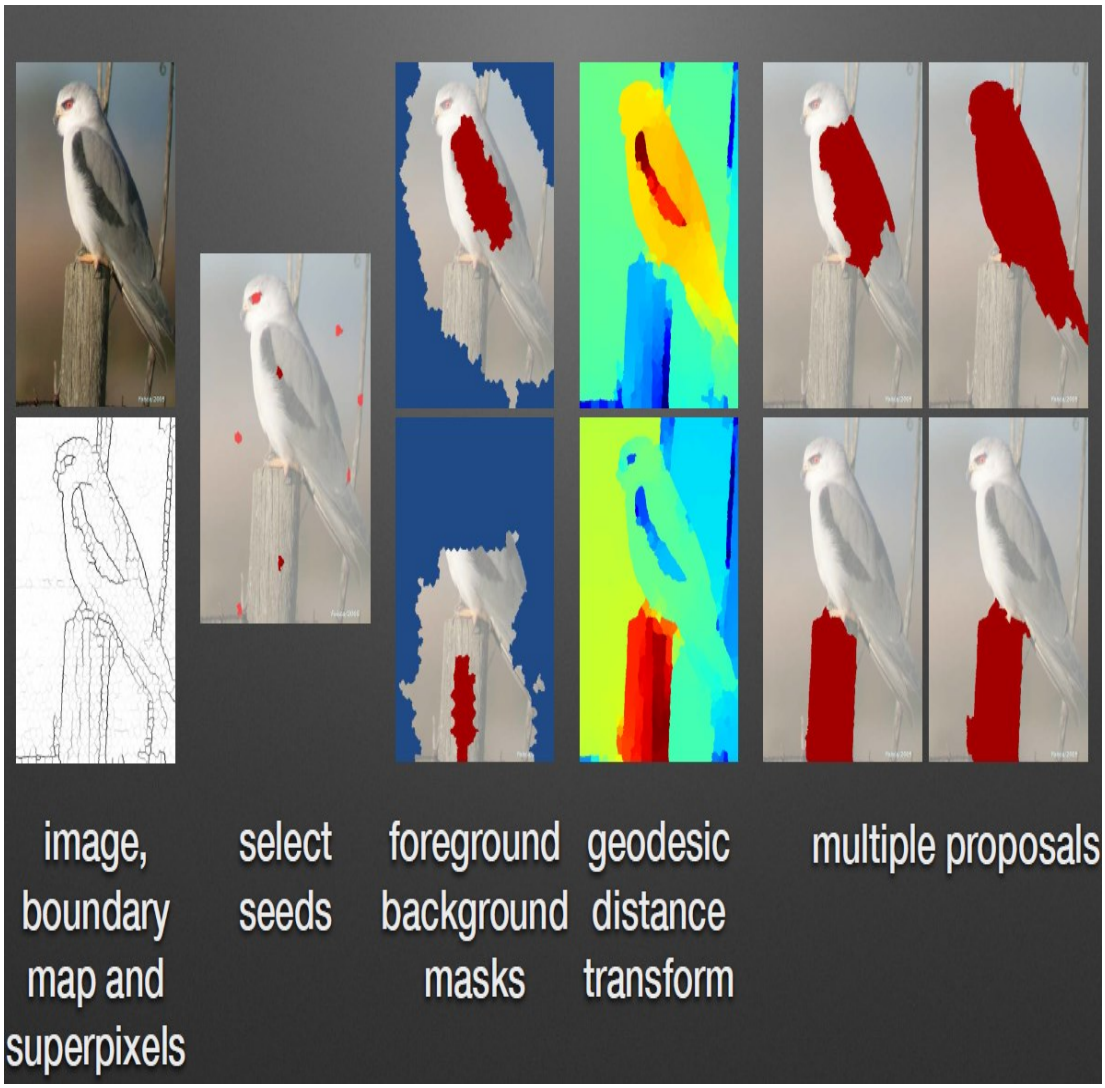

Figure 5a: GOP pipeline from left to right. Proposals are obtained by placing seed on the image. Eventually computing geodesic distance using seeds. Image courtesy of Krähenbühl.

3.2 Multiscale Combinatorial grouping

Pablo Arbelaez et al[14] suggested a composite method for image segmentation and object candidate generation, in their work Multiscale Combinatorial Grouping. The idea was to exploit multiscale hierarchal information for image segmentation. Then using smart combining technique, which would combine regions from different scales into possible object candidates. They addressed two topics. One was hierarchical segmentation and the second one was object segmentation proposals. This work has achieved some state-of-the-art results on different benchmarks.

The proposed methodology uses a bottoms-up hierarchical image segmentation technique. Initially a fast normalized cut algorithm is introduced that speeds up considerably the computation of Eigen vectors for contour globalization. As stated by the authors the prime difference between MCG and existing approach is, MCG is a unified approach that generates and then group together high quality multiscale regions. It does not depends upon pre-computed hierarchies and superpixels.

Image is segmented independently into multiple resolutions. See figure 5b. Forming an image pyramid. Each image now represents a family of super pixels, from fine set of superpixel to the complete domain. Each level of superpixel set is represented by a tree diagram which represents hierarchy of categories based on the fact that how similar or dissimilar they are tree structure is called a dendogram. Hierarchical representation of the image boundaries are called Ultrametric Contour Map UCM, further represents these sets by assigning the weight to the boundary of adjacent regions in the hierarchy by the index at which they were merged. Now simply, thresholding at a particular level in UCM produces segmentation.

In the second phase all hierarchical boundaries are aligned and combined in a multiscale hierarchy. Eventually a grouping component scan efficiently through the combinatorial spaces and produces a ranked list of object candidates. Ranking strategy utilizes the information about size, location, shape and contours.

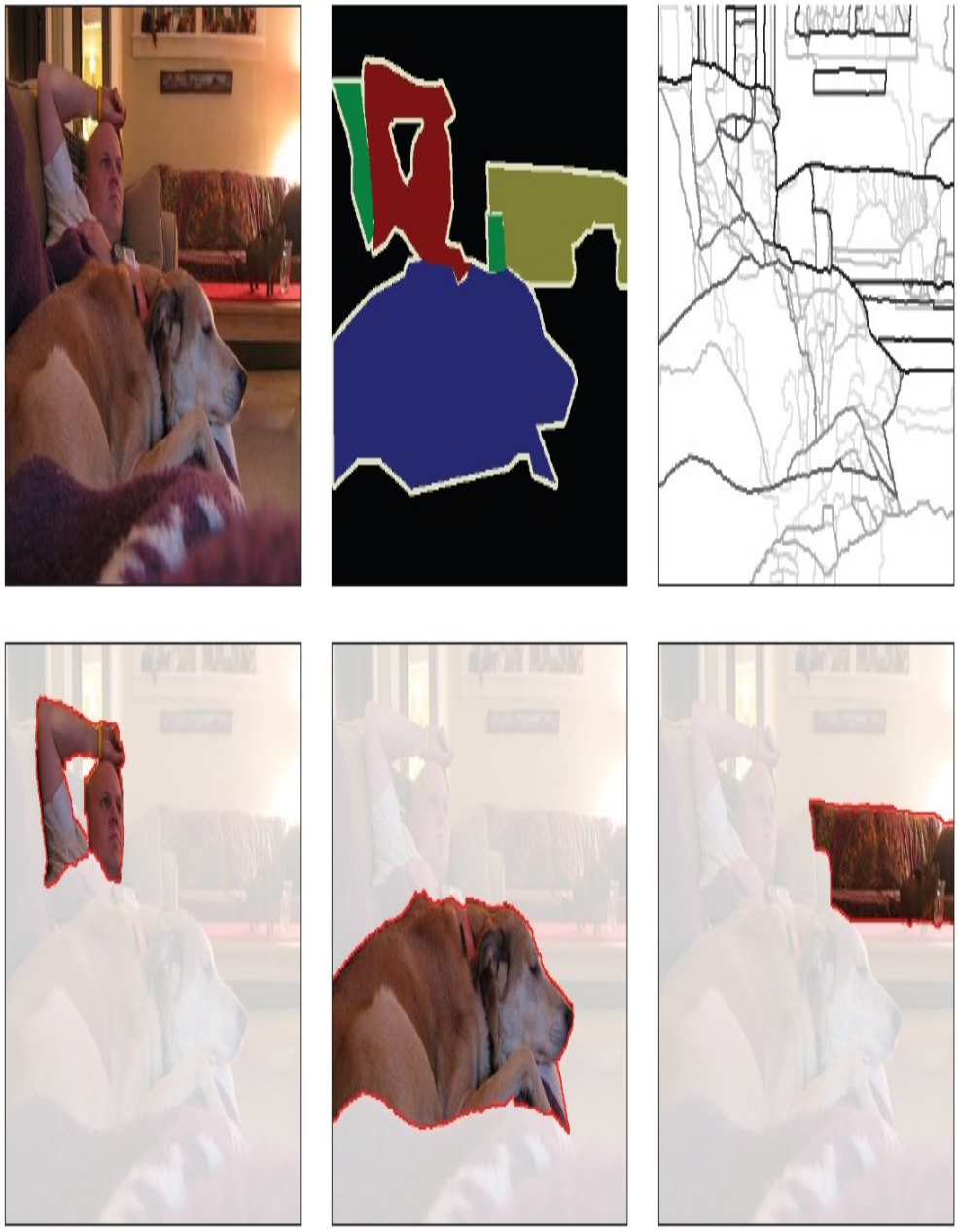

Figure 5b: Examples of MCG regions from [14].

3.3 Proposals with Global and Local Search

Proposals with Global and Local Search was proposed by Rantalankilla et al[13]. The main objective of their work was, to propose a method that produces accurate class-agnostic segmentations. These segmentations are then used as input to image recognition pipelines, reducing the computational cost. Their results on state-of-art dataset shows promising improvements both in terms of recall and computational time.

Proposals with Global and Local Search method can be distributed in to three steps. Oversegmentation of the input image is obtained by two methods named SLIC [36] and FH [37]. SLIC produces compact superpixels with roughly equal size whereas in contrast to SLIC, FH produces a diverse set of superpixlels which could range from half of the image to a very small object boundary. For every superpixel, feature vectors are computed using SIFT and RGB values are extracted from each pixel. Vectors are quantized using a learned visual vocabulary. Superpixels are refined based on similarity scores computed. The methodology then merges two similar superpixels into a larger superpixel and scores are updated. The algorithm is run till a specific similarity threshold is reached. Superpixels are refined by this process and superpixels from the previous stages are discarded.

In the second step superpixels are merged using a local approach. This approach considers only superpixels pair at once. Based on the visual similarity, score is assigned to each superpixel. Most similar superpixels are then merged and weights are updated accordingly. The process is iterative and runs till only one superpixel is left. All of the generated segmentation proposals are collected. Since it considers only two superpixels at once, part of an object that is similar to the background could get merged before the object is detected. Therefore, local approach is not recommended for large non-homogeneous objects.

Finally, in the global search all superpixels are considered. The problem is solved by computing optimization function over a graph. Nodes represent superpixels and edges represent the relation between adjacent superpixels. For every superpixel two labels are defined, background and foreground. Along with a unary term, a pairwise term is derived from the scores of adjacent superpixels. These parameters (unary,pairwise) along with the “label” hypothesis generates pool of segmentation proposals. Figure 6 shows superpixels generated by the discussed technique.

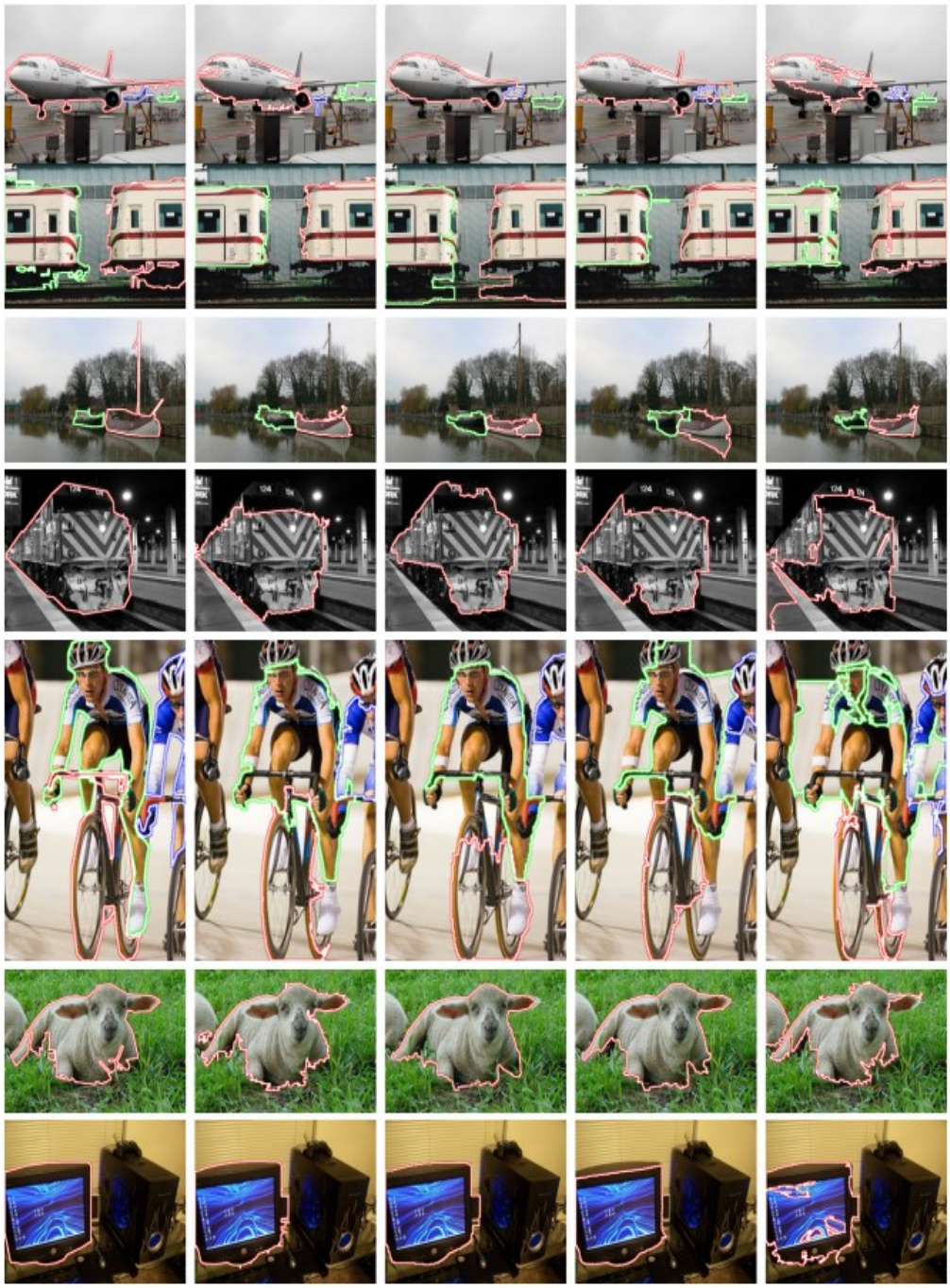

Figure 6: Examples of PGLS [5].

3.4 Selective Search

Selective Search for Object Recognition is one of the top performing algorithms on state-of-the-art benchmarks [6]. Proposed by Uijlings et al[8].The main aim of their work was to compose a method that would utilize segmentation and exhaustive search to generate accurate proposals.

In their paper they uses bottoms-up hierarchical segmentation. In a nut-shell they generate small segmentation proposals on all possible scales. For initial segmentation the given algorithm uses [37]. In the second step, iteratively initial segmentations are greedily merged together. The merging technique is based on similarity score. After combining two similar regions, similarity score is updated. The process continues till the whole image becomes one region. Similarity is measured primarily based on two features, Texture and Region.

Lastly, in order to diversify the set all initial segmentations with different starting parameters and on different color spaces are combined to generate class independent object segmentation proposals. It is one of the top performing algorithms in PASCAL VOC benchmark. Figure 7 sows the pipeline and hierarchy of the selective search method.

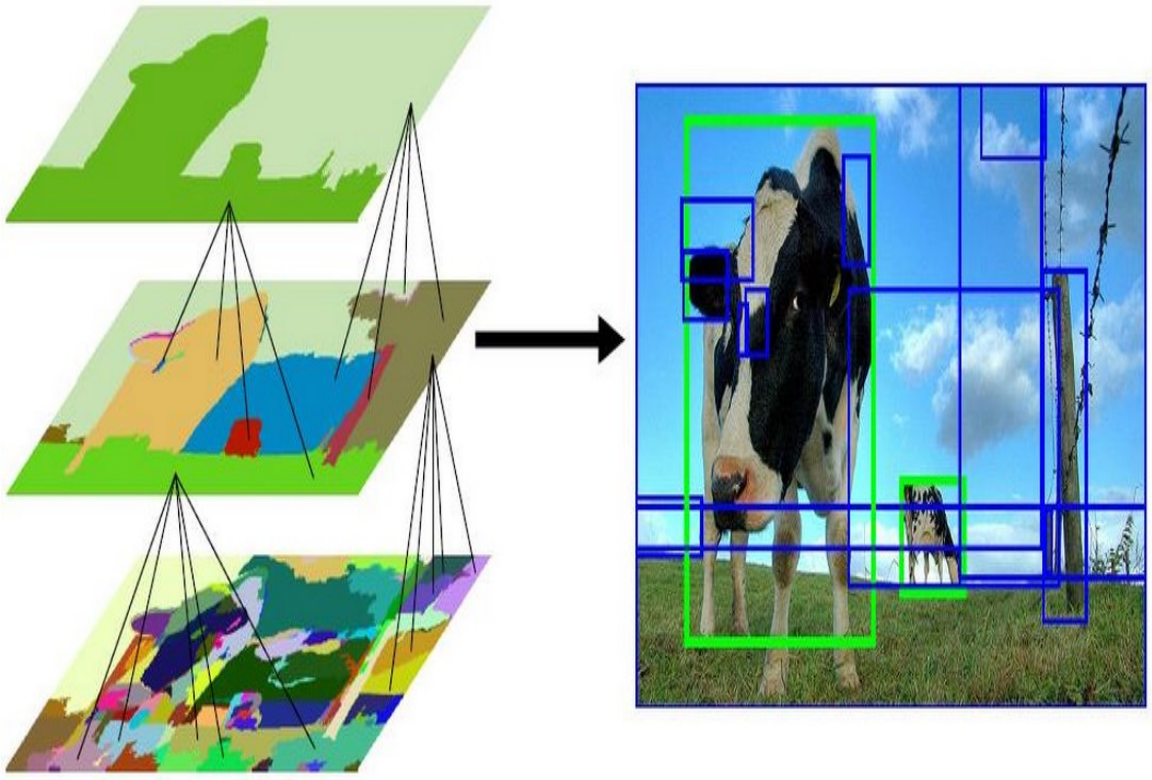

Figure 7: Selective Search Pipeline. Illustration adapted from [8].

3.5 Past Evaluation of Object Segmentations Proposals

In the recent past, J hosang et al[29] al published their work focused on the evaluation of object segmentation proposals. They conducted an extensive evaluation of twelve candidate generation methods(including these four) along with four baselines. They expanded their work from PASCAL VOC dataset to ILSVRC set. The principal goal of their research was to compare the existing methodologies in a standardize format. In their work, they alter the images by introducing different effects to the image such as JPEG artifacts, illumination variations, salt and pepper noise and rotation. Their work the evaluated the performance of candidate generation methods on these altered images and called this notion "repeatability". The motivation behind the work was to analyze if a method could produce the proposal for roughly same image content repeatedly. As discussed in the section 2.1.3, few current state-of-the-art methods deploy segmentation proposal method as preprocessing step for object detection. This research also conducted the effect of a different proposal method on R-CNN. They also proposed a novel evaluation metric AR (Average Recall) which was reported to correlates with detectors performance.

3.6 Summary

These four algorithms have obtained state-of-the-art results on current benchmarks. It is important to discuss the results of these methodologies, since they have reported differently in their papers. In terms of recall, MCG detects the maximum ground truth object at any overlap threshold. GOP produces the least amount of objects. However, all the methodologies were not tested on the same class or same model of the PASCAL VOC dataset. Besides recall, the computational time is another significant measure. Table 2 represents the running time of these algorithms along with average number of regions produced per image and mean average best overlap. As discussed above [29], have defined a new evaluation matrix for different proposal generation methods. Despite doing an extensive work on evaluation of these methodologies, they have adopted a different approach. They bring in some changes to the physical appearances of the images such as JPEG artifacts, illumination and rotation etc. Then they evaluate each method and obtain recall as a function of intersection over union on these images. Whereas our research focuses on to evaluate all of these four methods on single dataset without making any changes to the images. Then computing recall and average number of regions per image. This would give an idea of the performance of these methodologies and one can see which algorithm performs better than the other given same benchmark. Secondly, we also investigated how the strengths of these algorithms can be used together to further improve the performance. We started by combining the proposals of different algorithms together and eventually moved on to combine three of fastest methods. Experimental evaluation revealed that this improves the recall significantly even on challenging thresholds.

Table 2: Displays the computational time and number of regions produced using default settings on PASCAL VOC 2012 dataset.

Method	Time	# of regions per image	Average best overlap(bbox)
MCG	30s	5153	0.920
SCG	5s	2123	0.8905
PGLS	9s	1656	0.9120
Selective Search	4s	936	0.8641
GOP	8s	776	0.8859

4. EXPERIMENTAL EVALUATION

4.1 Segmentation Proposal Recall

To evaluate the quality of a segmentation proposal method, this research focused on the attribute that how well a given method detects ground truth object. Since, the objects that are missed will not be detected at all. In order to get more insights, our work conducted an in depth analysis of these four methods in terms of recall and further view the objects that are detected by a particular algorithm and not by others. This analysis further led us to the point where we were able to examine the weak links in these methodologies and finally, this apple-to-apple comparison led us to suggest some improvements in recall using these existing methods.

4.2 Evaluation Protocol

All experiments were tested on the validation set of PASCAL VOC 2012 challenge (1449 images). Objects that were labelled "difficult" in the dataset were also included in our experiments. To provide an un-biased evaluation, for all the methods default settings were used. However, for some methods additional segmentation proposals can be obtained by changing the parameters.

Jacard index was used to evaluate the quality of segmentation proposals. Jaccard index or Jaccard similarity coefficient is defined as intersection over union. Jaccard index of two regions A and B can be obtained by Equation 1**.** The value retrieved by the equation above is between the interval [0,1]. Value 1 depicts the complete overlap match between two regions and vice versa. The same evaluation criteria is used for bounding boxes as well, imagining bounding boxes as a rectangular region composed of pixels. The ground truth object is said to be detected, if and only if: the overlap score between the generated detection proposal and the ground truth object is greater than or equal to the overlap threshold defined.

The idea for obtaining recall for any of the algorithm was intuitively simple. After defining a minimum overlap threshold, count the number of objects for which the Jaccard similarity coefficient exceeds the minimum overlap threshold divided by the total number of ground truth objects. Two overlap threshold were used in our experiments [0.5 and 0.7] both for regions and bounding boxes. Finally, the curves were plotted as function of recall and average number of generated regions per image.

4.3 Results at Instance Level

After running the experiments, recall and average number of regions were drawn. Figure 8 and Figure 8a shows the comparison of recall between the methods at bounding box overlap threshold 0.5 and 0.7 respectively. It is evident from the figures that on a relatively lenient threshold (0.5) most methods perform well. However, when the overlap threshold was set to be more challenging (0.7), some methods such as GOP and SCG suffers a significant loss in recall whereas PGLS shows relative improvement.

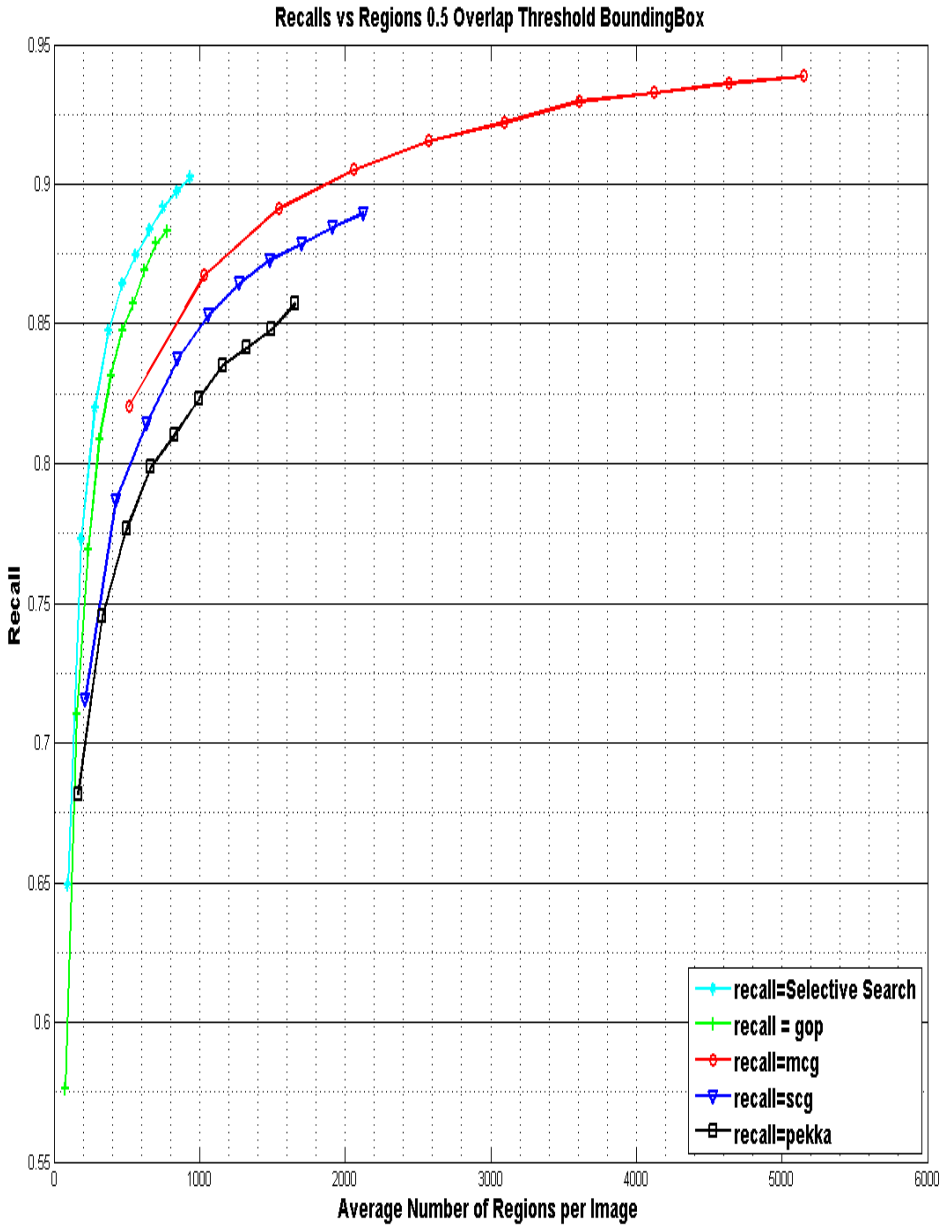

Figure 8: Recalls at 0.5 overlap threshold (bounding box). MCG is the best in terms of recall. Whereas GOP and Selective-Search are efficient in terms of regions generated.

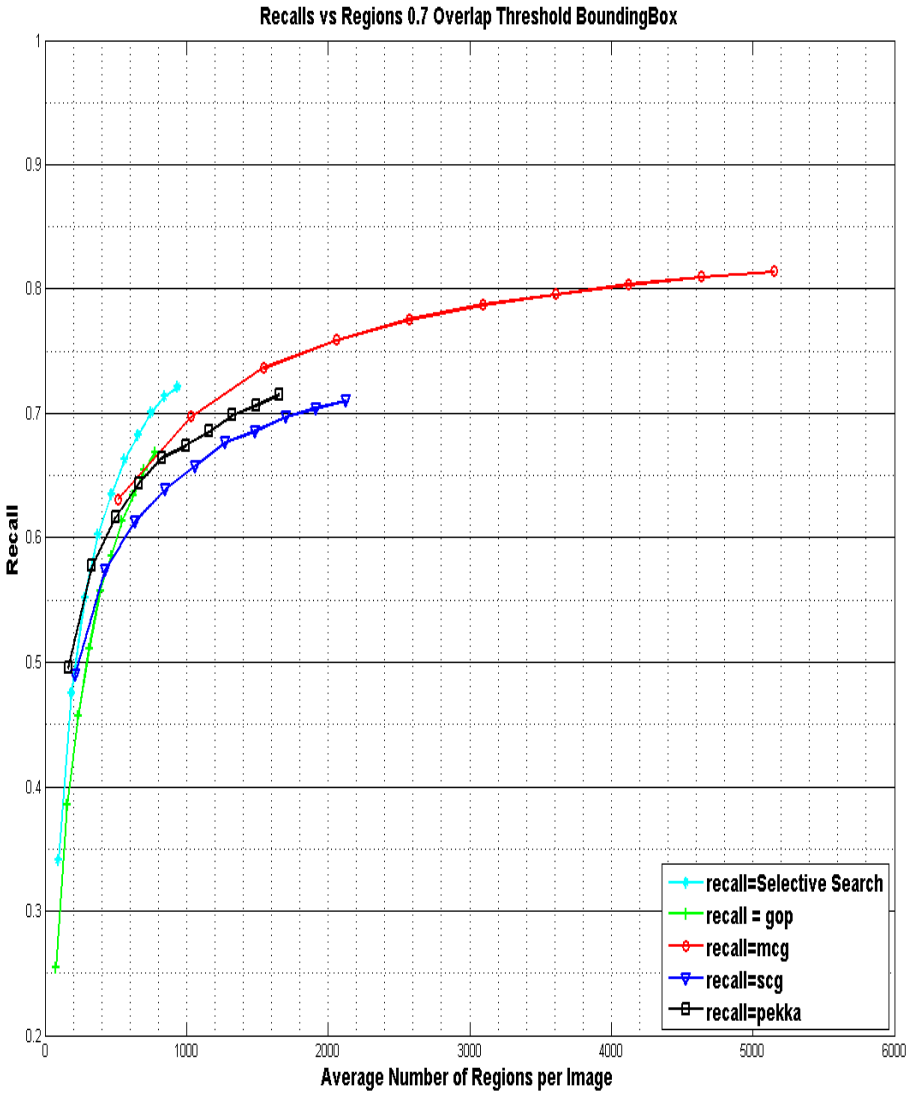

Figure 8a: Recalls at a more challenging 0.7 overlap threshold (bounding box). MCG is the best in terms of recall. Interestingly, PGLS here refer as pekka recall goes up from SCG at 0.7 threshold. GOP also suffers loss in recall.

Bounding Box overlap score can sometimes give inaccurate results. The actual region overlap could be small whereas the Jaccard Index for bounding box can return a high overlap score, this problem was also discussed in [12]. In order to avoid the overlap bias, Jaccard similarity coefficient was also computed for regions mask as well. Figure 9 and Figure 10 shows overlap for region masks and it is clear that all methods that perform well on bounding box threshold have less recall when region mask overlap was computed.

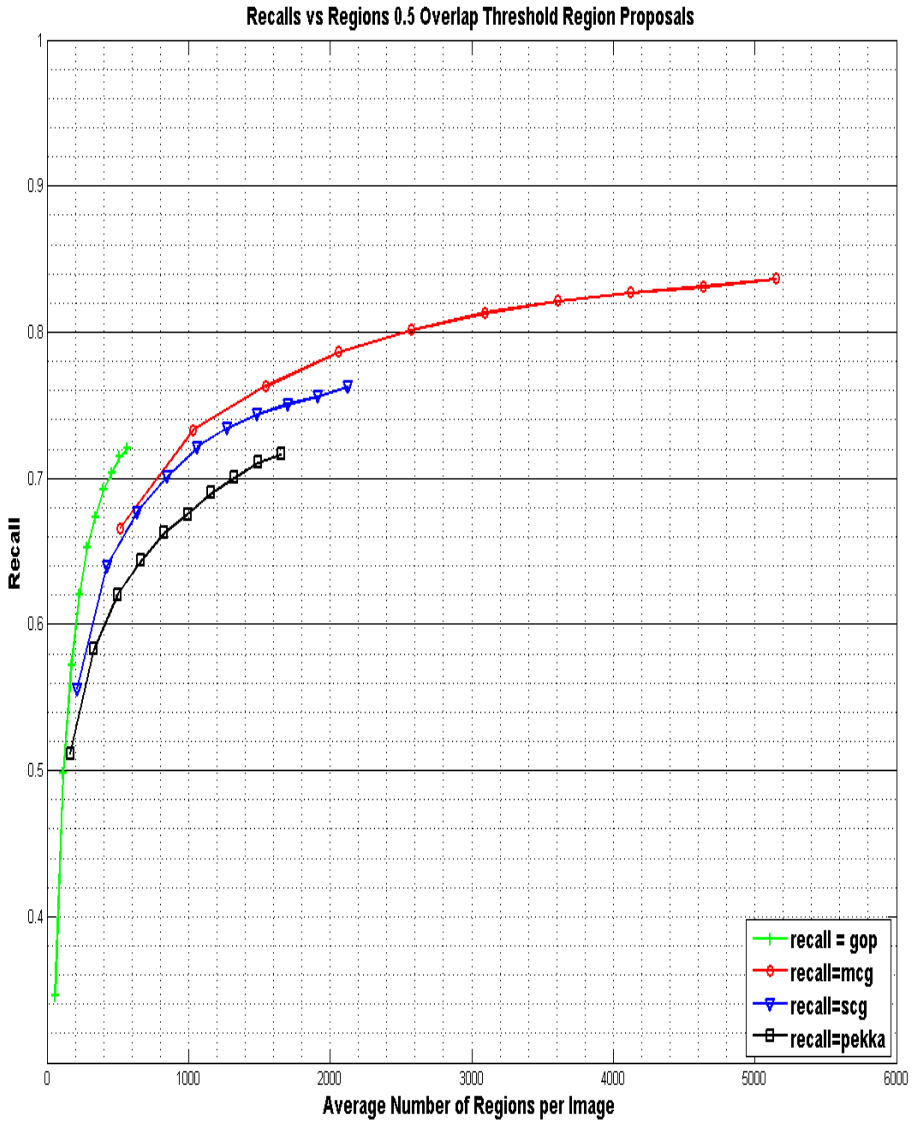

Figure 9: Recalls 0.5 overlap threshold (segmentation regions). MCG is the best in terms of recall.

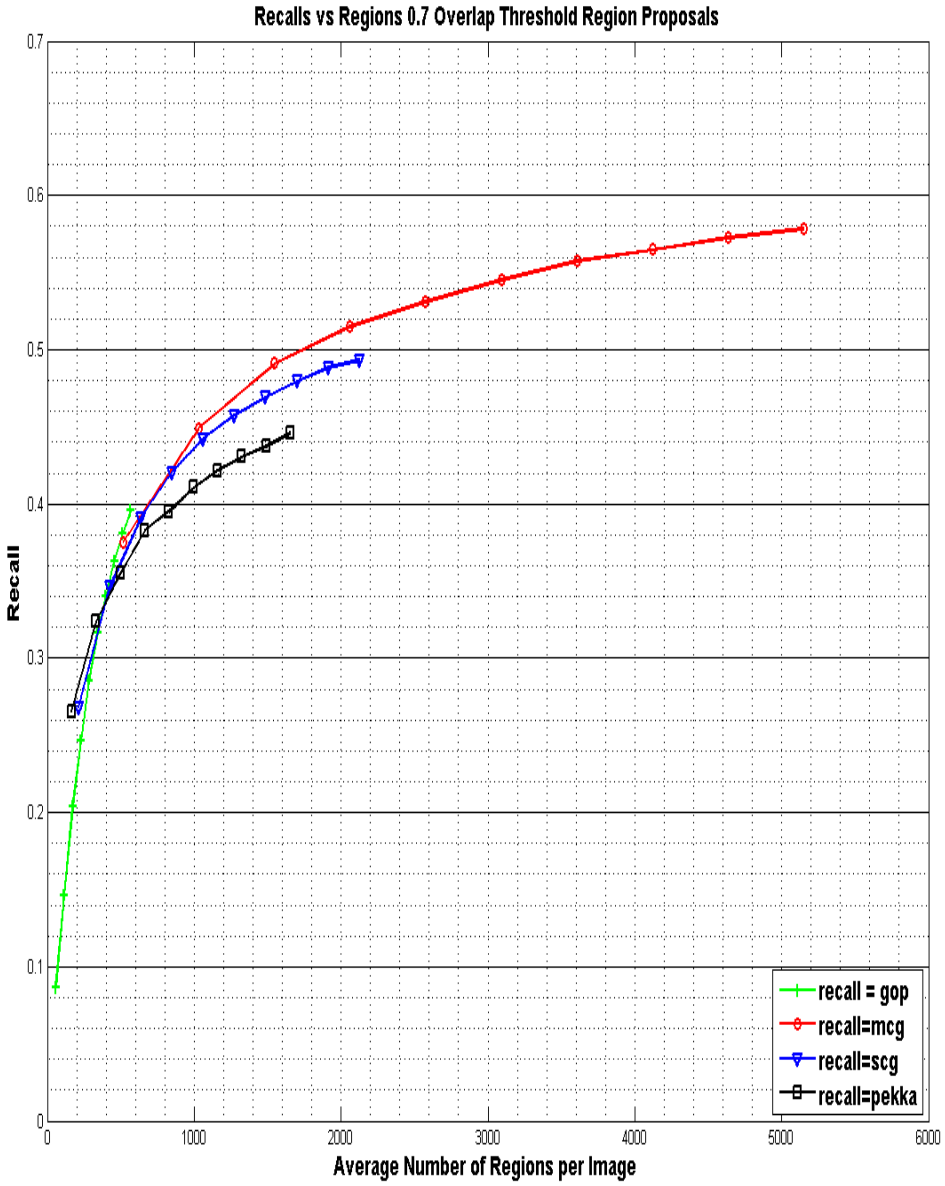

Figure 10: Recalls at a more challenging 0.7 overlap threshold (segmentation regions).Selective Search does not output segmentation masks using default settings. MCG has the highest recall.

It is evident that even for a less challenging overlap threshold for region masks, the recall for all the methods are below 0.85.It can again be observed that when these methods are compared to each other MCG performs best in terms of recall. However, it also produces large number of regions and hence is relatively computationally expensive. It was concluded from these experiments that 0.7 is an accurate overlap threshold in comparison with 0.5. Further in experimental evaluation we investigated the strength and weakness of these four proposal methods on different object classes.

4.4 Results at Class Level

To gain further insights, all four segmentation proposal methods were also evaluated in-terms of recall at class level. PASCAL VOC 2012 has 20 object classes. It was important to report the results, since all these methods present their results at class level in a different manner. Secondly, Karanbuhl and Koltun[12] did not report their results at class level. Results at class level also provides useful insights on, how a given method perform on objects of different shapes and sizes. Figure 11 shows the graph of all methods at class level.

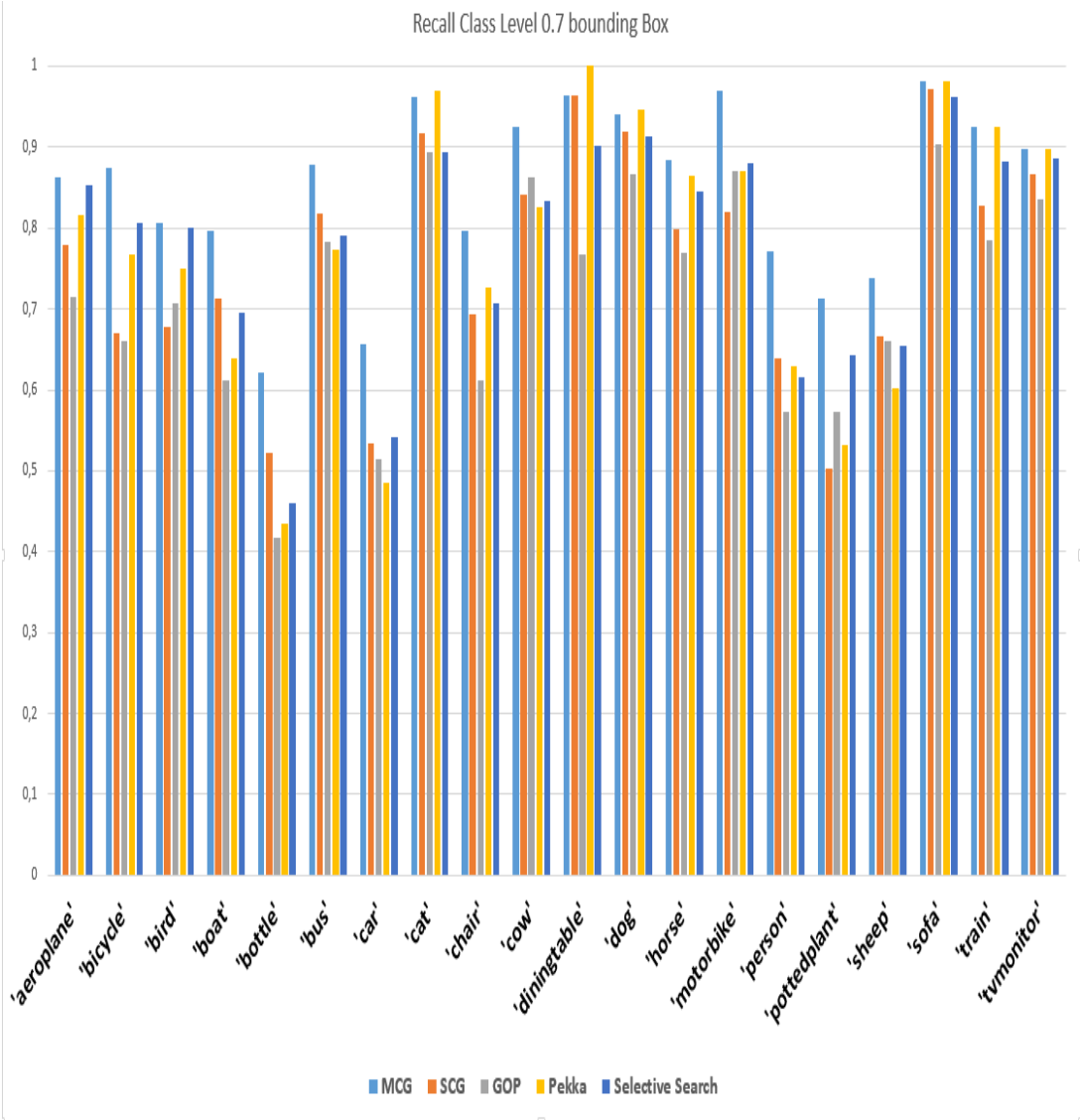

Figure 11: Recall values for different methods at bounding-box overlap threshold 0.7.

One could observe that “bottle” is a class where all algorithms perform poorly. Additionally, it can be seen that generally there is no clear winner at class level i-e either

all algorithms perform reasonably better or struggle on a particular class. Classes that occur naturally in a rectangular shape such as ‘dining table’ and ‘sofa’ are almost invariably detected precisely by all methods. In fact, on ‘dining table’ all ground truth objects were accurately detected by PGLS. Conversely it can be deduced from the Figure 11, all methods struggle on elongated objects such as ‘bottle’, ’potter plant’.

It was a stated fact that can be corroborated from our findings, one of the known drawbacks of GOP are small objects. Whereas comparatively, MCG performs better on small objects. However, if the overall performance is evaluated all algorithms roughly perform in the similar way on each object classes, keeping in mind the fact that how differently these algorithms are designed originally. This led us to the final development where we merge the algorithms together to gain possible improvements in recall. While merging, an important factor is the number of regions produced by each methodology. GOP produces the least number of regions, so even if it is combine with MCG the total number of regions would still be comparatively low.

4.5 Results with Combination of Different Algorithms

The main motivation behind this work was, combining the strengths of different algorithms together such that the average number of regions are still relatively less and recall is high. Figure 12 and Figure 13 presents the recalls of algorithms merged together, at both 0.5 and 0.7 bounding box overlap threshold.

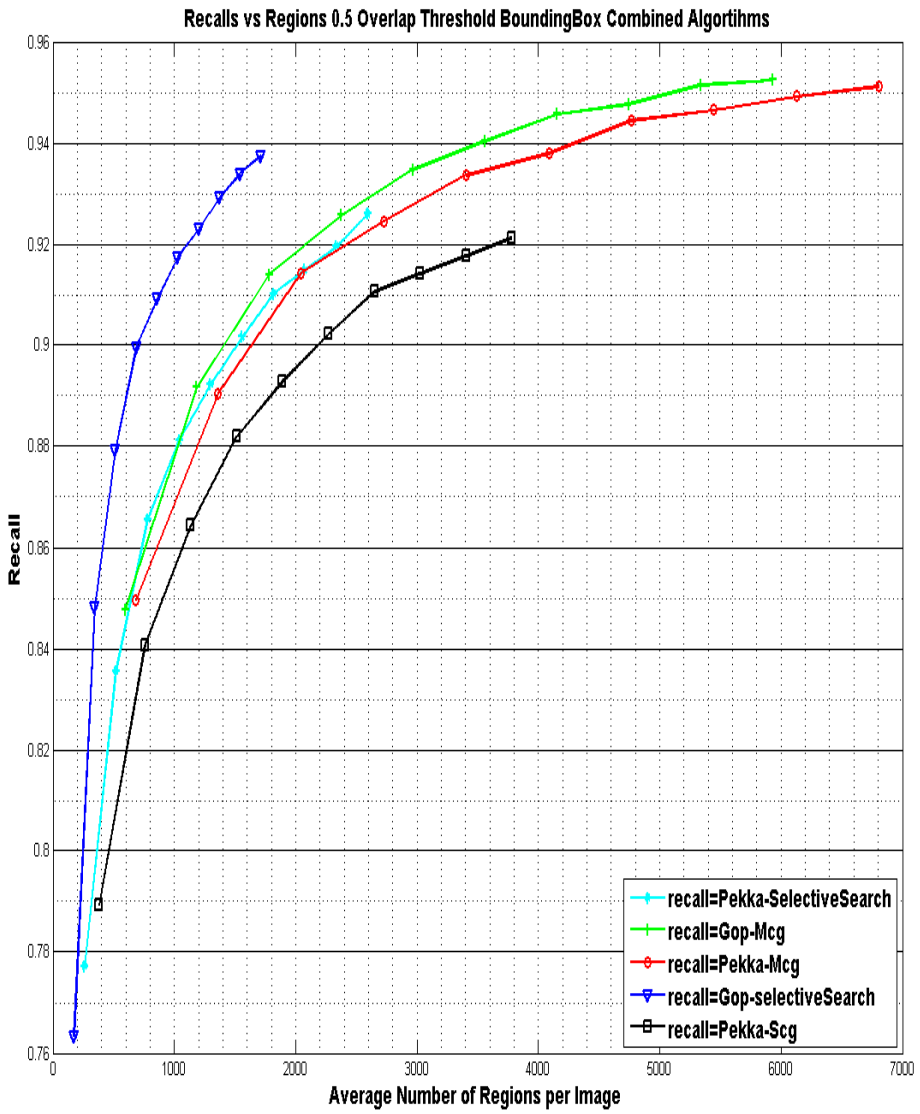

Figure12: Recall of merged algorithms at bounding box overlap threshold 0.5. Merger of GOP-MCG pushes recall to above 0.95 with less than thousand more regions. Whereas individually GOP had recall of less than 0.9, and MCG 0.95.

The clear improvement in terms of recall can be observed from Figure 12. Also, when GOP and Selective Search were combined the joint recall of the methodologies were almost same as of MCG for less than half average number of regions produced.

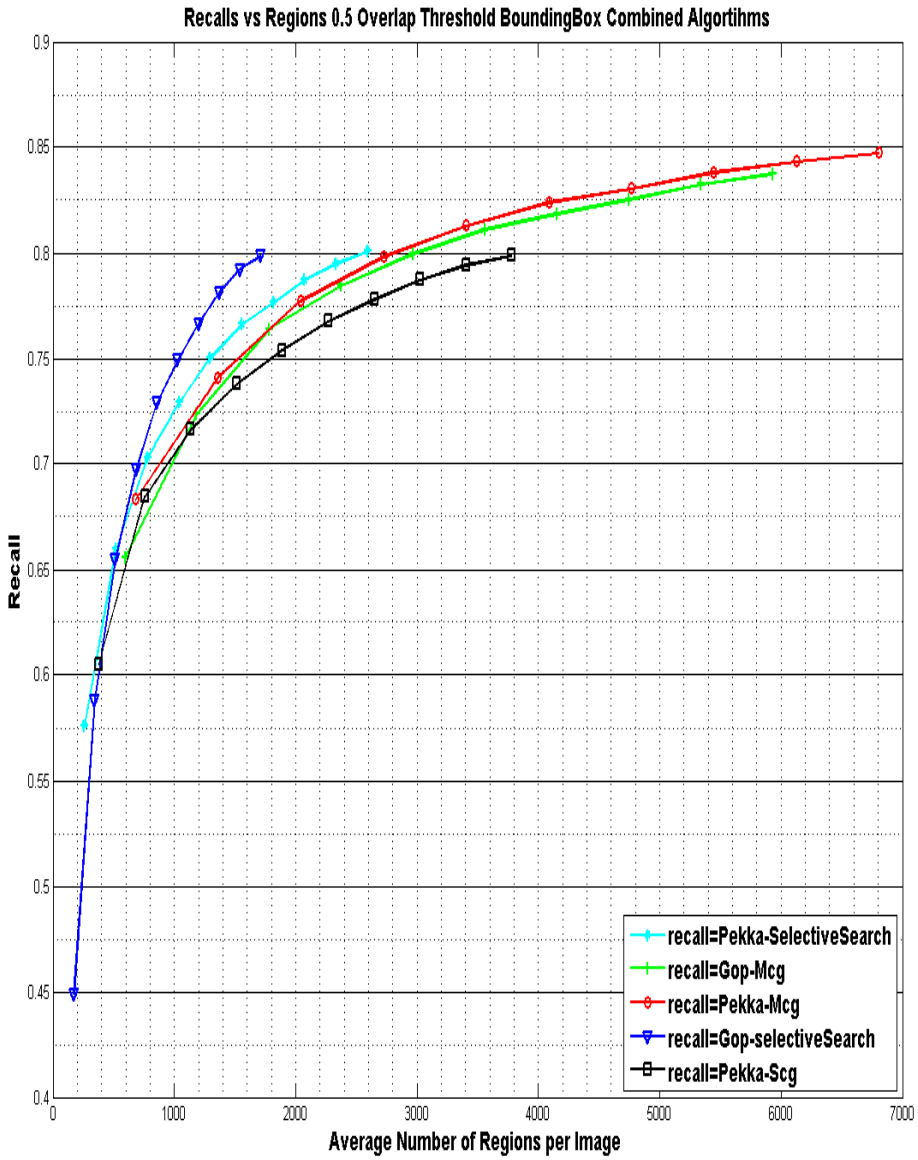

Figure 13: Recall of merged algorithms at bounding box overlap threshold 0.7.

It is clear from the findings that at a challenging threshold some algorithms suffer a loss in recall. However, a little increase in number of regions as compared to MCG the combination of MCG and GOP pushes the recall significantly up. This could be stepping stone in the object detection pipeline. Recall of almost 0.85 at this challenging threshold can be obtained by merging PGLS and MCG together. Although, they produce around 7000 average number of regions but it is still very less compared to the sliding window approach which produces around 10^4 to 10^5 number of windows per image. Like previously, results of some combinations were also evaluated at class level. Figure 14 presents the comparison graph of different combinations of algorithms together and individual algorithms at class level.

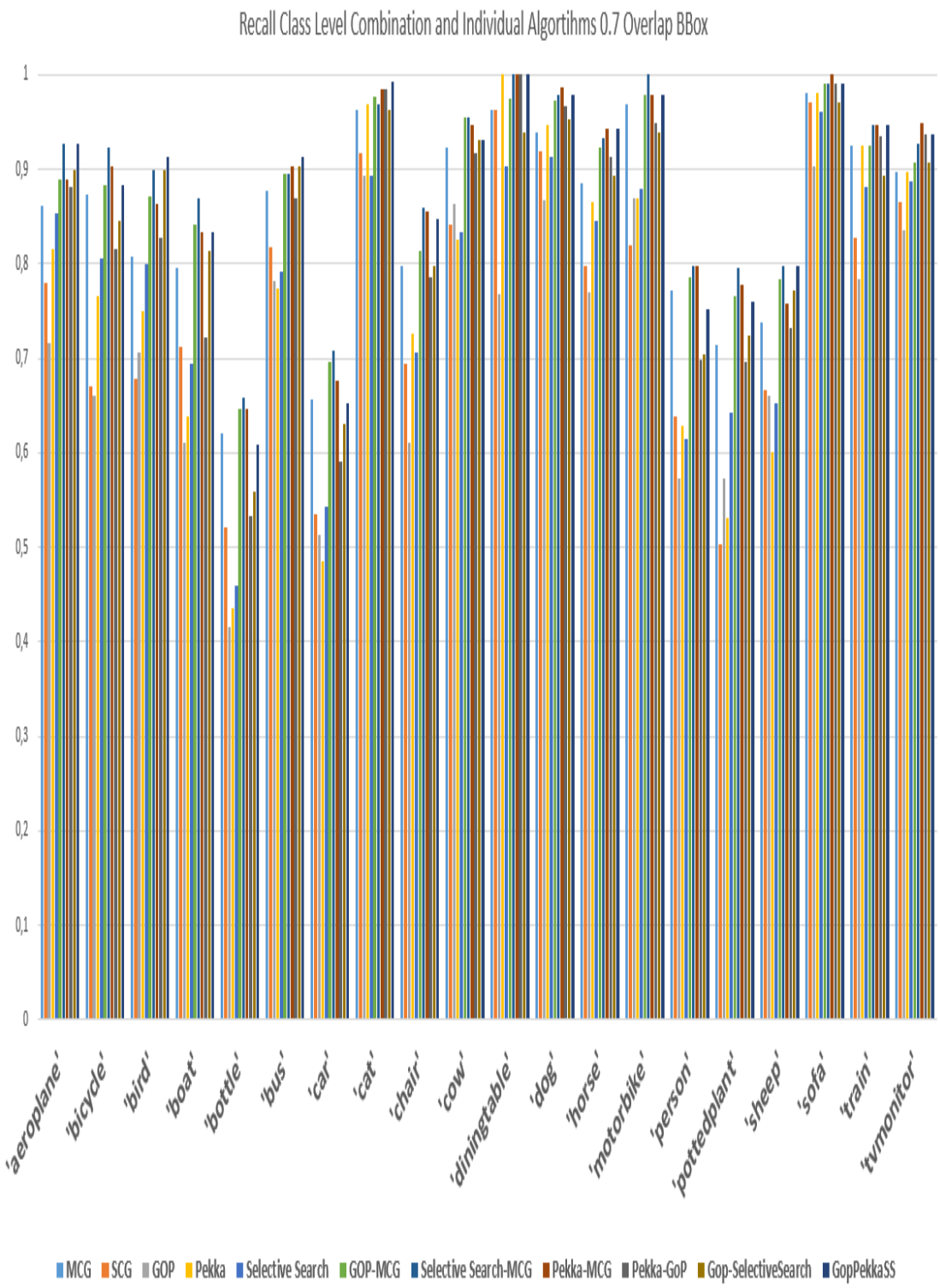

Figure 14: Class level representation of different algorithms and their combinations.

Improvements can be seen in every class. Although, like previously no combination can be regarded as a clear winner at class level. One interesting finding is that for elongated objects for which previously all individual method suffered low performance such as “Bottle”,”Potter Plant” the combination of MCG-Selective Search per-

forms better than all other algorithms and combination tried. These results further supports the initial hypothesis that combination of algorithms would further increase the recall at fractional increase in computational time. These findings led us to one step further, we merge three of the fastest methods, Selective Search-PGLS-GOP. Figure 15 represents the recall of this triplet at 0.7 overlap threshold.

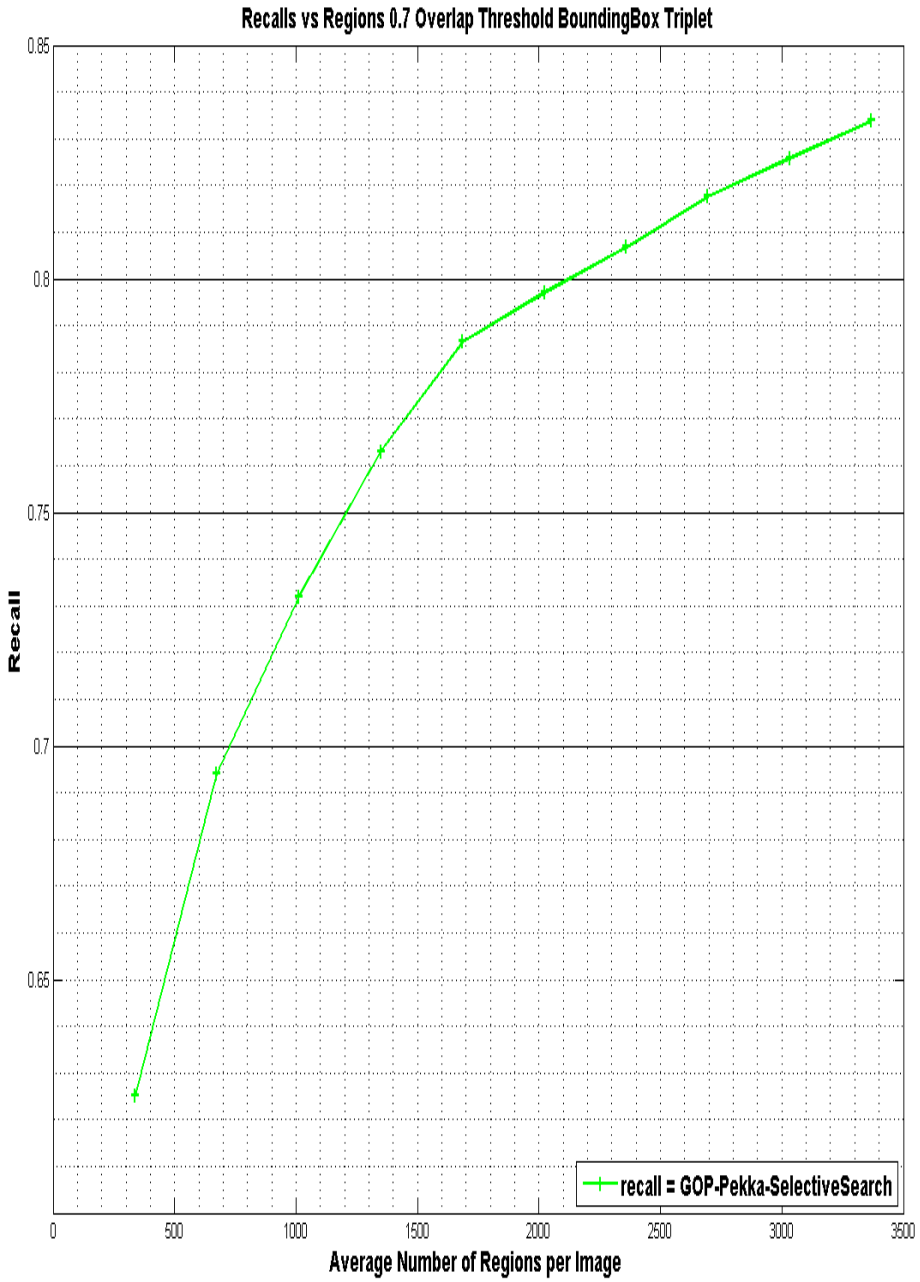

Figure 15: Recall of three algorithms together (SS-PGLS-GOP). It has significantly higher recall than MCG(0.81). It produces just more than half the number of regions than MCG.

Although, higher recall can be achieved by merging PGLS and MCG but it is computationally expensive, since the combination produces around 7000 average number of regions whereas this triplet produces almost half number of regions, recall is slightly lower but importantly it is two times faster than the combination. We achieved the highest recall at a challenging Jaccard Index(0.7) by combining the proposals of all the algorithms together. Figure 15a shows the comparison between triplet and combination of all methodologies together. Clearly, the combination of all algorithms pushes the recall significantly further up, one needs to also observe the number of regions as well as it crosses the mark ten thousand proposals.

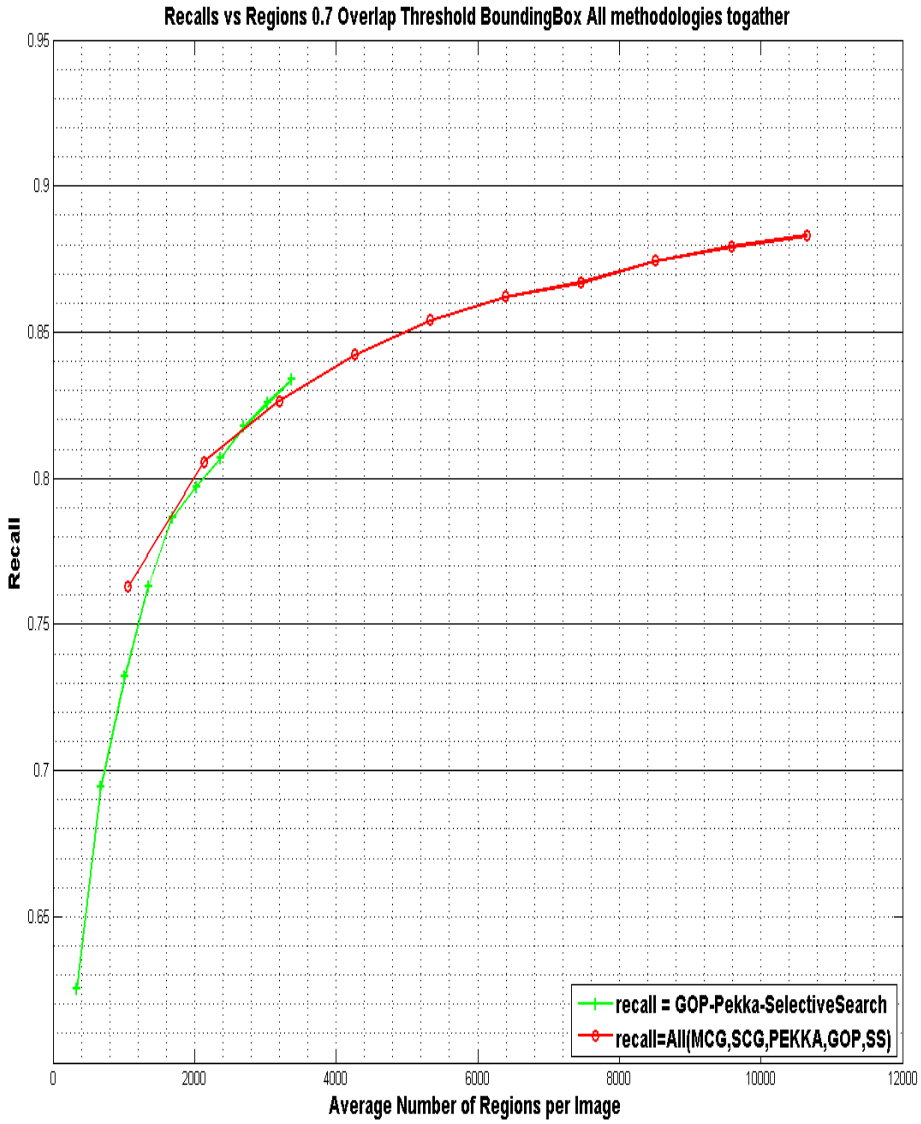

Figure 15a: Comparison of recalls between the triplet and the combination of all the proposals at 0.7 overlap threshold. The highest recall can be observed by combining all proposals. However, this also produces largest number of regions.

Another test that was not done previously by the authors was, testing of the algorithms on objects labeled as “Truncated” and “Difficult” on PASCAL VOC 2012.As previously explained in Figure 3b as well,”Truncated” object were defined as objects which extend beyond the bounding box. Object annotated as “Difficult” were objects which were not scored in evaluation. Besides these two, there were two other annotations per object, Pose and Occluded. They are outside the scope of this thesis and therefore not included in our research.

All algorithms were tested on objects labelled as **Difficult** and **Truncated** in PASCAL VOC 2012 validation set. Figure 16 shows the distribution of ground truth objects along with individual algorithms, triplets and this time we also combined all four algorithms together. However, it was also revealed that combining all four algorithms increases computational time with a very small improvement in recall.

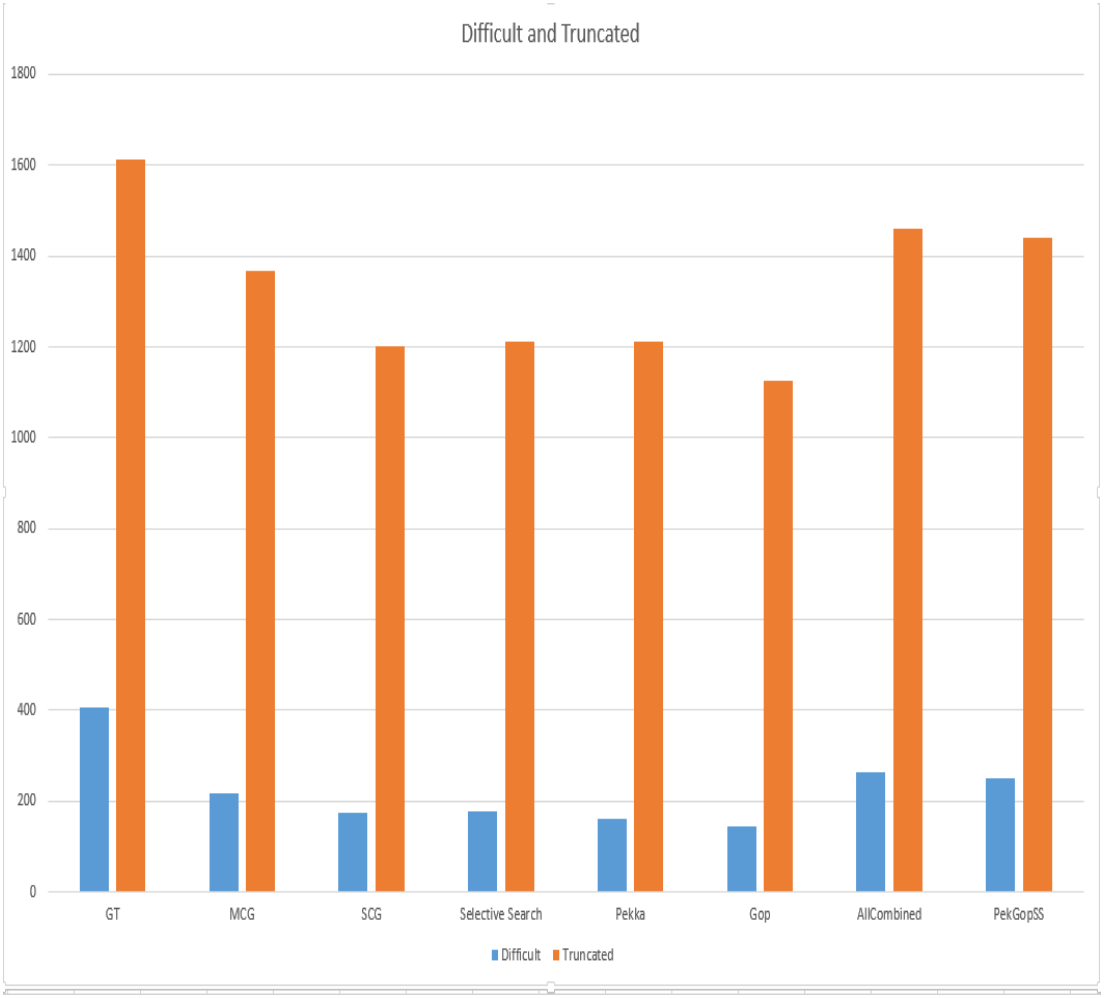

Figure 16: GT bars shows the total number of Difficult and Truncated objects in the dataset (validation). Each bar above an algorithm depicts the detected objects by the algorithm at 0.7 bonding box overlap threshold. It can be seen that all algorithms combined pushes the recall higher than triplet of (PGLS-GOP-SS) but the difference is quite small compare to the computational cost.

From Figure 16, it can be concluded that apart from MCG all individual algorithms failed to detect more than half of the objects annotated “Difficult”. Whereas the triplet performs better than MCG as well and as discussed previously it is quicker and produces less number of regions than MCG. Keeping in mind the objects annotated “Difficult” are not completely present in the image scene, it would not be fair to penalize algorithms based on this conclusion only. However, it was still important to test it on “Difficult” objects, since it tells us the expected performance of these algorithms on more practical and challenging images. As opposed to “Difficult” the “Truncated” objects were detected relatively efficiently by all algorithms. Contrary to “Difficult”, “Truncated” objects were present in the scene and only small part of the object was truncated. Therefore “Truncated” objects were relatively not as challenging as “Difficult” objects are.

In order to gain further insights of the algorithm’s strength and weaknesses toward particular type of objects. Two more experiments were conducted to further investigate the physical nature of the objects that are detected and missed by these individual algorithms and the triplet that we have formed. Figure 17 and 18 shows the distribution of size in pixels and aspect ratio of the ground truth objects and objects detected by the methodologies at 0.7 bounding box overlap threshold respectively.

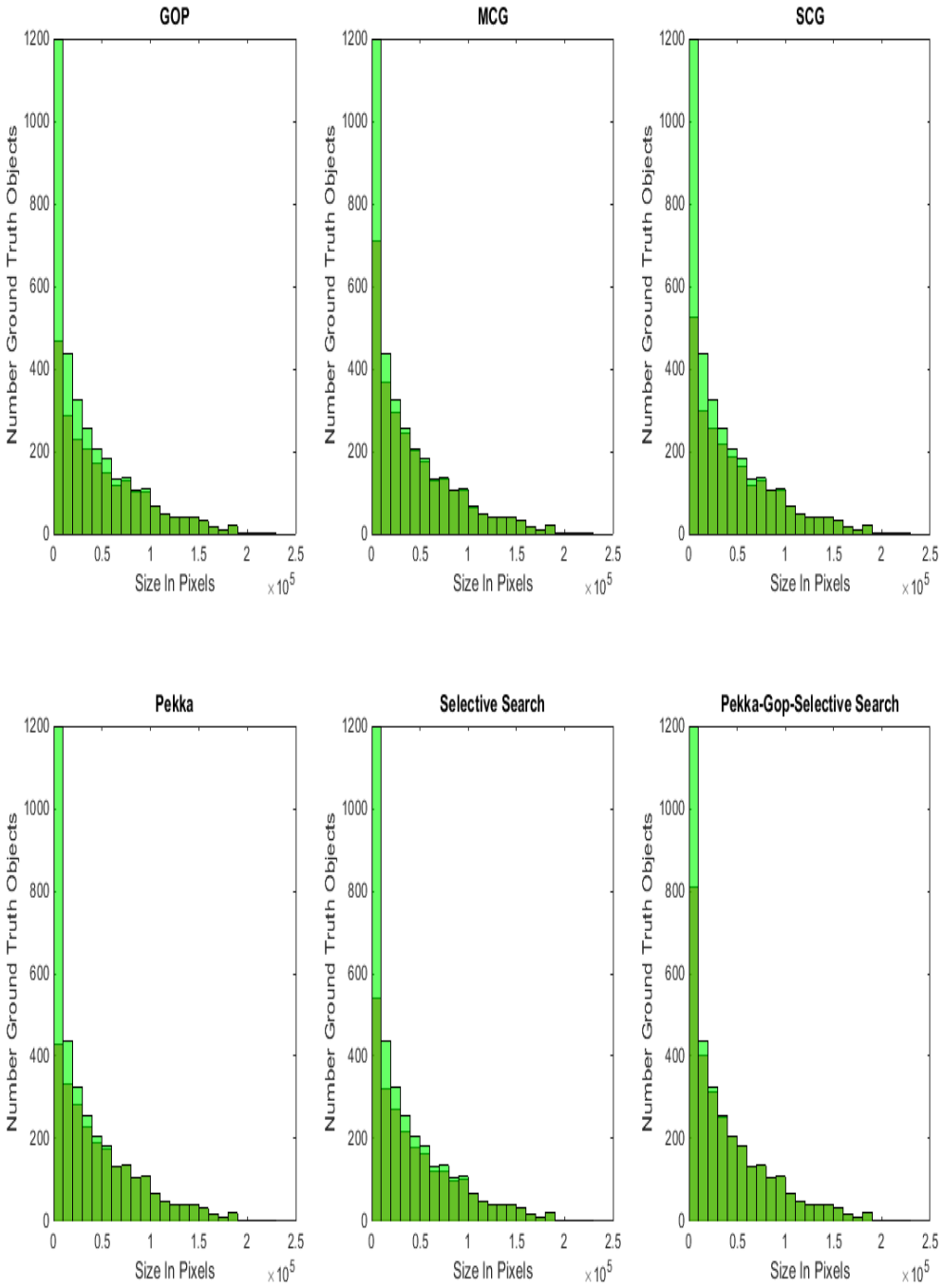

Figure 17: Distribution of objects in terms of size in pixels of the Dataset. The length of bars represent total number of ground truth objects of that particular size. Filled area shows the amount detected by the corresponding algorithm.

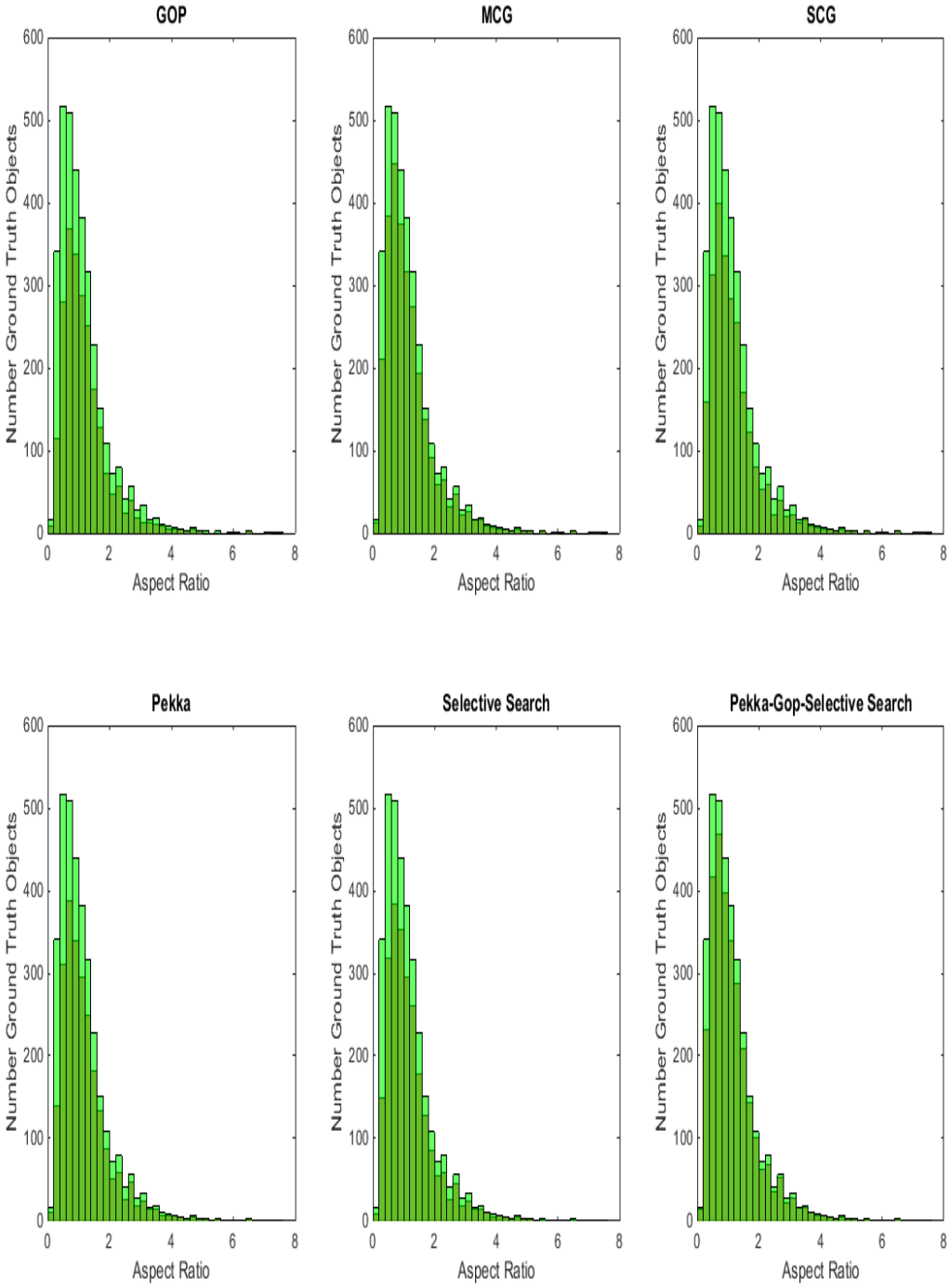

Figure 18: Distribution of objects in terms of aspect ratio of the Dataset. The length of bars represent total number of ground truth objects of that particular aspect ratio. Filled area shows the amount detected by the corresponding algorithm.

Finally, Figure 17 and 18 further reveals the fact, segmentation proposals methods perform relatively better on large objects and objects that naturally occur in rectangular shapes I-e objects whose horizontal dimension is greater than its vertical one. Objects such as “airplane”, “motorbike” and “sofa” are some of the objects that are less challenging for current state-of-the-art methods. Importantly, these experiments again show that algorithms struggle on elongated and tall objects. Interestingly, all algorithms relatively perform poorly on small objects and the combination of three quickest method improves the recall at each class and object level. This further support the fact that these algorithms have similar strength and weaknesses. While in future designing new methods for object detection proposals these strength and weaknesses should be considered.

5. DISCUSSION AND CONCLUSION

In this thesis we presented an evaluation of four recent algorithms in segmentation proposals. During the course of this thesis we also provided some background knowledge of object detection and how it has shaped over the years to better understand the background of the problem. The paramount goal of this study was to make a comparison between different segmentation proposal methods. While conducting the research, we narrowed it to the different object classes where we were able to identify the objects on which these methods perform fairly well and where these methods suffered low accuracy. This aspect could aid future researchers to choose a candidate generation method specifically for those object categories in which they are interested in. We used Jaccard index to compute accuracy of different methodologies. We also concluded that 0.5 overlap threshold seemed to be inaccurate and 0.7 is a more reliable threshold and most experiments were conducted using this higher threshold.

Experiments revealed that MCG is the best performing methodology, has the highest recall and is suitable for most object classes. MCG outperforms all algorithms on maximum object classes of PASCAL VOC dataset. However, MCG is computationally expensive and produces the largest number of regions. A variant of MCG, SCG which uses single scale instead to multiscale lags behind PGLS and Selective Search when overlap threshold was set to be more challenging. If number of regions is the main concern, GOP produces the least number of regions but its recall is comparatively low. GOP suffers low recall on small objects. However, GOP is one of the fastest method to date. Selective Search produces fairly accurate proposals and is robust. Selective Search for example generates less than half number of regions than SCG, still it comes second in almost all object classes losing only to MCG that produces roughly five times more regions than Selective Search. Besides MCG, Selective Search is the best method over all for Non-Rigid objects. PGLS method achieves highest recall for several object categories including “Cat” “Dining table” “Dog” ”Sofa” ”Train” and “TV Monitor”.

We proposed and showed improvement in recall by combining these existing methods. We started by merging two algorithms that produce less regions, primarily keeping the computational overhead in mind. Pairing of algorithms proved to produce improvements in recall as compared to individual algorithms. However, to gain insights we also combined proposals of different algorithms with MCG but the improvement in recall was relatively low. We made a triplet by merging three of the fastest methods PGLS, Selective Search, GOP. Experimental evaluation revealed that it out performs the top performing MCG in terms of recall and yet produces almost half the number of regions as MCG produces. This indeed could be a new avenue to explore as to the best of our knowledge no previous works have explored this aspect.

Interestingly, despite being dissimilar in the nature these algorithms operate, they share generally the same strength and weaknesses. The object categories that have good recall are common among algorithms where they perform better and vice versa. The current four approaches work well on the object class with large regions, they all suffer a lot when the objects are small and elongated. It can also be observed, that the algorithm that exploits multi scale or operates on multiple color channels seems to have a higher recall than other approaches.

In future, we would like to evaluate the performance of R-CNN and similar systems if they are built on top of the triplet that we have formed. Secondly, we would also like

to address the issue of low performance on certain elongated objects. We plan to make public all the segmentation masks, bounding boxes generated by these method, in standardize format and the scripts used for evaluation along with this document.

6. REFERENCES

[1] Kootstra, G., Bergstrom, N., & Kragic, D. (2010, December). Fast and automatic detection and segmentation of unknown objects. In Humanoid Robots (Humanoids), 2010 10th IEEE-RAS International Conference on (pp. 442-447). IEEE.

[2] Shotton, J., Blake, A., & Cipolla, R. (2005, October). Contour-based learning for object detection. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 1, pp. 503-510). IEEE.

[3] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on (Vol. 1, pp. 886-893). IEEE.

[4] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on (Vol. 1, pp. I-511). IEEE.

[5] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR 2004, Workshop on Generative-Model Based Vision. 2004.

[6] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2), 303-338.

[7] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014, June). Rich feature hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 580-587). IEEE.

[8] Uijlings, J. R., van de Sande, K. E., Gevers, T., & Smeulders, A. W. (2013). Selective search for object recognition. International journal of computer vision, 104(2), 154-171.

[9] Wang, X., Yang, M., Zhu, S., & Lin, Y. (2013, December). Regionlets for generic object detection. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 17-24). IEEE.

[10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575.

[11] Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In Computer Vision–ECCV 2014 (pp. 391-405). Springer International Publishing.

[12] Krähenbühl, P., & Koltun, V. (2014). Geodesic object proposals. In Computer Vision–ECCV 2014 (pp. 725-739). Springer International Publishing.

[13] Rantalankila, P., Kannala, J., & Rahtu, E. (2014, June). Generating object segmentation proposals using global and local search. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 2417-2424). IEEE.

[14] Arbelaez, P., Pont-Tuset, J., Barron, J., Marques, F., & Malik, J. (2014, June). Multiscale combinatorial grouping. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 328-335). IEEE

[15] Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons.

[16] Agarwal, S., & Roth, D. (2002). Learning a sparse representation for object detection. In Computer Vision—ECCV 2002 (pp. 113-127). Springer Berlin Heidelberg.

[17] Ponce, J., Berg, T. L., Everingham, M., Forsyth, D. A., Hebert, M., Lazebnik, S & Zisserman, A. (2006). Dataset issues in object recognition. In Toward category-level object recognition (pp. 29-48). Springer Berlin Heidelberg.

[18] Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2014). The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1), 98-136.

[19] Mundy, J. L. (2006). Object recognition in the geometric era: A retrospective. In Toward category-level object recognition (pp. 3-28). Springer Berlin Heidelberg.

[20] Dickinson, S. J. (2009). Challenge of image abstraction. Object categorization: computer and human vision perspectives, 1.

[21] R. Brooks. Model-based 3-D interpretations of 2-D images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 5(2):140–150, 1983.

[22] Lowe, D. G. (1984). Perceptual organization and visual recognition (No. STAN-CS-84-1020). STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

[23] Fergus, R., Perona, P., & Zisserman, A. (2007). Weakly supervised scale-invariant learning of models for visual recognition. International Journal of Computer Vision, 71(3), 273-303.

[24] Manen, S., Guillaumin, M., & Gool, L. V. (2013, December). Prime Object Proposals with Randomized Prim's Algorithm. In Computer Vision (ICCV), 2013 IEEE International Conference on (pp. 2536-2543). IEEE.

[25] Carreira, J., & Sminchisescu, C. (2012). Cpmc: Automatic object segmentation using constrained parametric min-cuts. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(7), 1312-1328.

[26] Alexe, B., Deselaers, T., & Ferrari, V. (2010, June). What is an object?. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 73-80). IEEE.

[27] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32(9), 1627-1645.

[28] Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. IEEE Transactions on Computers, 22(1), 67-92.

[29] Hosang, J., Benenson, R., Dollár, P., & Schiele, B. (2015). What makes for effective detection proposals?. arXiv preprint arXiv:1502.05082.

[30] Hosang, J., Benenson, R., & Schiele, B. (2014). How Good are Detection Proposals, really?. In 25th British Machine Vision Conference (pp. 1-12). BMVA Press.

[31] Chang, K. Y., Liu, T. L., Chen, H. T., & Lai, S. H. (2011, November). Fusing generic objectness and visual saliency for salient object detection. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 914-921). IEEE.

[32] Humayun, A., Li, F., & Rehg, J. M. (2014, June). RIGOR: Reusing Inference in Graph Cuts for generating Object Regions. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 336-343). IEEE.

[33] Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In Computer Vision–ECCV 2014 (pp. 391-405). Springer International Publishing.

[34] Rahtu, E., Kannala, J., & Blaschko, M. (2011, November). Learning a category independent object detection cascade. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 1052-1059). IEEE.

[35] Cheng, M. M., Zhang, Z., Lin, W. Y., & Torr, P. (2014, June). BING: Binarized normed gradients for objectness estimation at 300fps. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on (pp. 3286-3293). IEEE.

[36] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 34(11), 2274-2282.

[37] Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. International Journal of Computer Vision, 59(2), 167-181.