TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

JUNAID MALIK

CATEGORY INDEPENDENT OBJECT PROPOSALS USING QUANTUM SUPERPOSITION

Master of Science thesis

Examiners:
Prof. Moncef Gabbouj
Dr. Çağlar Aytekin
Examiner and topic approved by the Faculty Council of the Faculty of Computing and Electrical Engineering on
9th November 2016

ABSTRACT

JUNAID MALIK: Category Independent Object Proposals Using Quantum Superposition
Tampere University of technology
Master of Science Thesis, 51 pages, 0 Appendix pages
January 2017
Master's Degree Programme in Information Technology
Major: Signal Processing
Examiners: Prof. Moncef Gabbouj, Dr. Çağlar Aytekin

Keywords: object detection, object proposals, quantum cut quantum superposition, category-independent

A vast amount of digital images and videos are continually being generated and shared across the Internet. An important step towards utilizing this 'big data' and deducing meaningful information from its visual contents, is to detect the presence of objects belonging to a particular class in digital images. Earlier computer vision algorithms devised for this purpose exhaustively search the entire image space for detecting objects belonging to a particular class. Object proposals aim to reduce this search space by proposing probable locations of objects in the image beforehand. This paves the way for efficiently using more computationally expensive and sophisticated detection algorithms.

Conventional approaches to generating object proposals have revolved around learning a scoring function from the characteristics of objects in ground truth annotations of images. In this thesis, we propose a novel category independent proposal generation framework that is unsupervised and inspired by the psycho-visual analysis of human visual system where the search for objects gradually transitions from the most salient parts of a scene to comparatively non-salient regions. We use a state-of-the-art visual saliency estimation technique which proposes a unique relationship between spectral clustering and quantum mechanics. We improve this method by exploiting for the first time, the quantum superposition principle, to extend the search of objects beyond the salient ones. We also propose an unsupervised scoring strategy that does not incorporate any prior information about the spatial, color or textural features of objects.

Experimental results have proved that our proposed methodology achieves comparable results with the contemporary state-of-the-art methods. Our unsupervised scoring strategy is shown to outperform, in some cases, the supervised frameworks employed by other methods. Moreover, it also enables us to achieve a three-fold decrease in the number of proposals while keeping the loss of recall to less than 3%. The success of our proposed methodology opens the door to a research direction where quantum mechanical principles can be utilized to enable computer vision algorithms to find objects in digital images without having any prior knowledge about them.

PREFACE

The research work presented in this thesis has been conducted at the Department of Signal Processing, Tampere University of Technology.

I would like to start by thanking my supervisor Prof. Moncef Gabbouj for providing me with the opportunity to be a part of the MUVIS research group and conducting the research work for this thesis. I would especially like to thank my examiner, Dr. Çağlar Aytekin for his continuous technical and motivational support throughout the thesis process.

Lastly, I would like to dedicate this thesis to my mother, Rukhsana Perveen and my siblings. None of this would be possible without their prayers and endless moral support.

Tampere, January 2017

Junaid Malik

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

IoU	Intersection over Union
QCUT	Quantum Cut
EQCUT	Extended Quantum Cut
MCG	Multiscale Combinatorial Grouping
ABO	Average Best Overlap
NMS	Non-Maximum Suppression
MABO	Maximum Average Best Overlap
CRF	Conditional Random Field
UCM	Ultrametric Contour Map
BING	Binarized Normed Gradients
CPMC	Constrained Parametric Min Cuts
MMR	Maximum Marginal Relevance
SIFT	Scale Invariant Feature Transform

1. INTRODUCTION

Advancements in semiconductor manufacturing has led to significant miniaturization of integrated circuits which has enabled wide-scale production of complicated electronic circuits and their usage in consumer electronics. Image acquisition devices are no exception to this. Once considered expensive and luxurious commodities, digital cameras are now essential to modern lifestyle by virtue of smartphones. This, coupled with the exponential rise in Internet speed, has ushered in an age of information explosion where digital visual content is being generated and consumed at a very high rate. In order to ensure efficient indexing and retrieval of this content, it becomes imperative to gain a scenic understanding of these images and videos. This problem has been at the center of computer vision based research and continues to be an active research area.

One of the key steps towards gaining an understanding of natural scenes is the detection and recognition of objects present in it. Human visual system is very adept at this task. Psycho-visual analysis has revealed that from a very small age, humans are able to detect contour segments and group them together based on their context [1][2] which forms an integral part of object localization, detection and eventually, recognition. Moreover, visual attention process is understood to consist of two sequential steps; a faster initial step which is task-independent and focuses on the most appealing parts of the scene followed by a later, comparatively slower step which searches the visual space for a particular object of interest [3]. Computer vision algorithms that aim to model the human visual system can be broadly divided into two categories along the same lines. Those which aim to model the former task-independent visual attention stage, termed as "visual saliency estimation methods", and the ones modelling the latter task-specific step of the process, commonly referred to as object detection and recognition methods.

Object recognition methods aim to solve the problem of finding if an object of a particular class is present in an image or not. Object detection on the other hand aims to simultaneously detect and localize i.e. define the spatial extent of an object. Localization is important for applications such as robotics where the precise location of an object is required in order to interact with it. The objects being searched for generally belong to a finite set of classes like faces [4], vehicles [5] and animals [6] etc. As there is no prior information available about the number and location of objects in the image, object detection methods search the entire image in an exhaustive manner where randomly sampled windows of arbitrary size and location are evaluated. This approach makes the process computationally expensive and prohibits use of more complex detection algorithms.

Object proposal generation aims to alleviate this issue by producing a list of probable locations of objects in the image. It acts as a precursor to object detection and aims to find multiple, possibly overlapping, regions in an image which have a high likelihood of covering the object(s) of interest. By focusing on a set of proposed regions, the computational load on the object detector being employed is reduced significantly as it doesn't have to traverse the search space of the whole image. Although an object detection or recognition scheme is always limited to recognizing objects belonging to a finite set of classes, object proposal generation methods are desired to be class-invariant to ensure that they are not biased towards any predefined set of classes.

In this work, we propose a novel approach to for generating category-independent object proposals. Taking inspiration from the way human visual system goes about the task, we start with the salient parts of the image and progressively expand our search of objects to other relatively non-salient parts of the image.

The rest of the thesis is structured as follows:

- Chapter 2 consists of a brief review of the state-of-the-art object proposal generation methods and a qualitative evaluation of the methodologies adopted followed by a brief introduction of the proposed methodology and the rationale behind it.
- Chapter 3 elucidates the proposed methodology for generating and ranking proposals.
- Chapter 4 benchmarks our method against the state-of-the-art by evaluating multiple performance metrics.
- Chapter 5 provides interpretation of the results, conclusions drawn from the work and recommendations for further research.

2. RELATED WORK

The design and evaluation of a category-independent proposal generation mechanism relies on a definition of an object in a digital image. Forsyth et al [7] provided a broad classification of the contents of digital images into *materials*; entities that have a specific texture or pattern but have no defined shape or size and *things*; those having a distinctive shape and well-defined spatial extent. Building upon these foundations, Alexe et al [8] further proposed that an *object* is a *thing* that has a well-defined boundary, uniqueness among its neighboring regions in the image and individuality in the context of the entire image. Object proposal generation methods aim to utilize one or more of these cues in the bid to propose regions of an image having a high likelihood of enclosing objects of interest.

The first work related to generating proposals was performed by Alexe et al [8]. Since then, a range of methods have been proposed which have taken different routes to generate category independent proposals [9]–[19]. Bounding box or window based methods [8], [18], [11], [19] evaluate rectangular windows sampled from the image based on a defined scoring function and the top scoring windows are proposed as detection proposals. Region based methods [9], [10], [12]–[17] (also referred to as grouping based methods in [20]) provide proposals in the form of pixel-wise segments, either by merging various over-segmentations of the image or by solving multiple foreground segmentation problems.

Following is a brief account of the proposal generation strategies adopted by the state-of-the-art methods in each category.

2.1 Window-based Methods

Window-based methods usually inherit the exhaustive search strategy of object detection methods where rectangular windows of different sizes, locations and aspect ratios are evaluated. However, instead of the comparatively complex task of detecting an object of a particular class in the window, the objective here is to identify windows which are most likely to contain an object. Figure 1 shows a generalized data flow of a window based proposal generation methods.

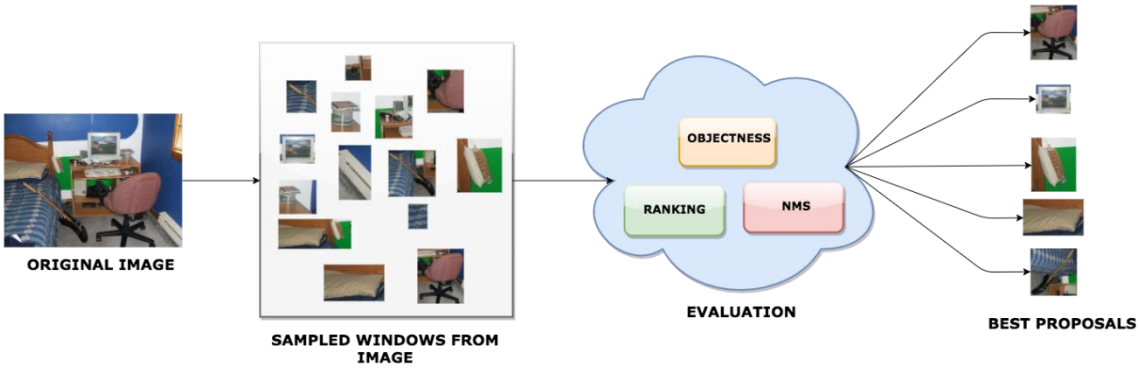

Figure 1. *Data-flow for bounding-box based methods: Windows sampled from the image are scored based on an objectness measure and ranked.*

An "objectness" measure is defined to quantify the likelihood of a window to contain an object. Figure 2 shows some examples of rectangular window candidates sampled from the images. An accurate objectness measure should assign a high score to windows colored green as they provide very good bounding boxes for the objects of interest. The windows colored blue should be penalized for not providing a tight enough fit and finally, the red windows should be scored very low as they do not enclose any object.

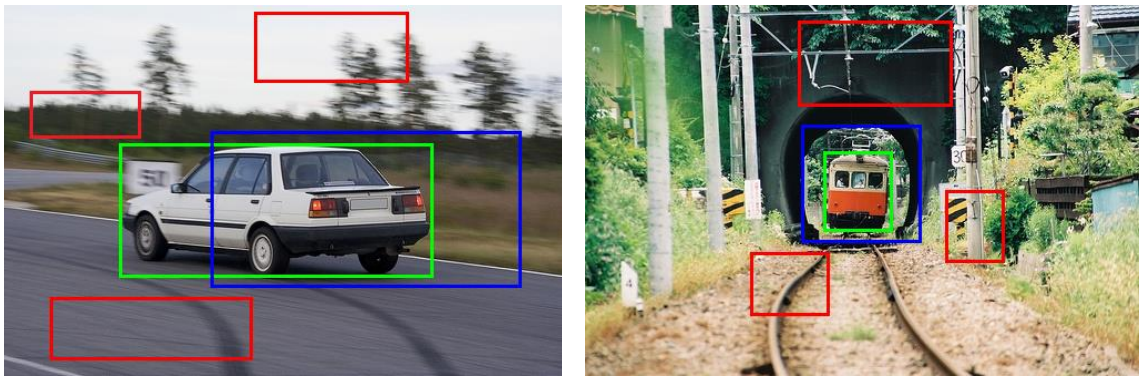

Figure 2. *Example of object windows providing precise localization of an object (green), partially covering the object (blue) and not enclosing any object (red)*

2.1.1 Objectness

The work done by Alexe et al [8] is one of the first in the field of object proposal generation. The authors leverage on a general fact that an object in an image has a well-delineated boundary and possesses a certain degree of uniqueness in its appearance, both from its immediate surroundings and globally within the image. Based on these, the authors train an objectness measure which evaluates rectangular windows in an image and scores them based on their likelihood of enclosing an object.

The objectness measure proposed by the authors comprises of three cues based on global characteristics of the image and local features of windows. The first cue, termed as *multiscale saliency (MS)* employs global saliency maps of the image calculated at multiple

scales using [21]. Saliency maps assign each pixel a score representing its prominence in the context of the image. The multiscale saliency score for a window is calculated by summing up the saliency values greater than a learnt threshold θ_{MS} and normalizing by the size of the window to remove the bias towards larger windows.

The second cue measures the amount of color disparity between a rectangular window and areas in its immediate proximity. This is done by increasing the size of the rectangular window in all directions by a factor θ_{CC}. The distinctiveness of the window is then calculated as the Chi-Square distance in the Lab color space between the original window and its enlarged version.

Figure 3. *Objectness: Best proposed window using* [8] *(red) and the ground truth (green)*

The next two cues exploit the closed boundary characteristics of objects in an image using two different techniques. First one, termed as *edge density*, is based on the concentration of edges around the border of a window. The window is first shrunk from all sides by a factor θ_{ED} . The border of the window is then defined as the inner ring between the original window and its contracted version. An edge map for the image is obtained using [22]. The density of the edges inside the border normalized by the parameter of the original window is then used as an objectness cue. The second cue is also loosely based on contours and is termed as *superpixel straddling*. It leverages on the fact that homogeneous oversegmentations of an image, or superpixels, preserve the object boundaries [8], [23]. A window is therefore penalized if it has a large number of superpixels *straddling* or cutting its boundary.

The values for the free parameters are learnt using annotated ground truth data. The optimal parameters for each of the cues is obtained by maximizing the posterior probability of correctly classifying object windows. The cues are then combined based on a Naïve Bayes framework, where they are assumed to be mutually independent. The final objectness score of a window is given as in (1). Any subset A of the set of all cues $\boldsymbol{C}$ can be used to calculate the objectness score.

$$p(obj|A) = \frac{p(obj)\prod_{cue\in A} p(cue|obj)}{\sum_{c\in\{obj,bg\}} p(c)\prod_{cue\in A} p(cue|c)} \quad (1)$$

Figure 3 shows an image taken from the PASCAL Visual Object Challenge (PASCAL VOC) [24] 2007 test dataset. The green window represents the ground truth and the red one represents the most accurate proposal produced by the objectness method.

2.1.2 Rahtu's Method

Rahtu et al [19] proposed a new method by revisiting the work done by [8] and building upon it by reducing the initial number of candidates and also improving the objectness measure by introducing two new cues for measuring objectness. Figure 4 shows the result of this method on two images taken from the PASCAL VOC [24] 2007 dataset.

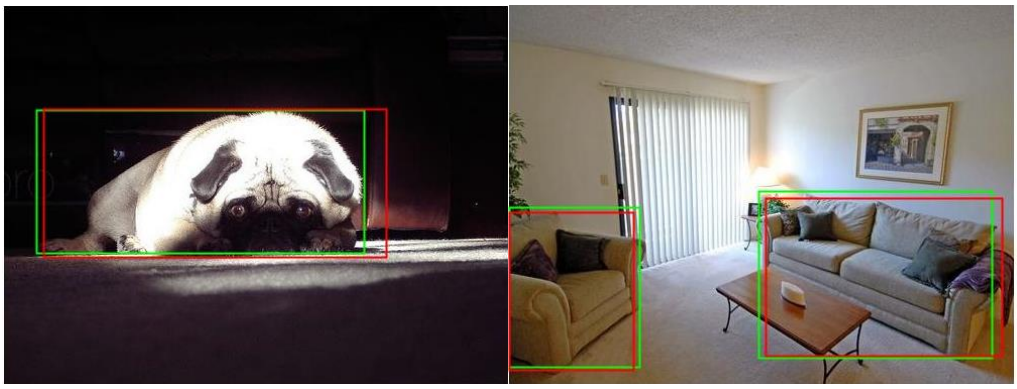

Figure 4. *Best windows proposed using Rahtu's method* [19] *(red) and ground truth (green) on two test images*

For a given image, the proposal generation algorithm starts by producing a set of initial rectangular windows. Instead of randomly sampling candidate windows from the image, the method uses two different ways to generate the initial pool of proposals. First is to use the bounding boxes that enclose each of the superpixels and connected pairs and triplets of them. This is based on the hypothesis that each superpixel belongs to a homogenous patch in the image and similar superpixels can be merged to possibly provide more accurate and complete boundaries of the objects. The second technique used to populate the initial set of windows relies on sampling candidates from a bounding box prior which corresponds to bounding box features (location and size) learnt on the ground truth annotations provided in the PASCAL VOC 2007 [25] training set. All the windows in the initial pool of proposals are then evaluated using an objectness measure.

The method proposes three new objectness cues which are similar to the ones proposed by Alexe et al in [8]. The *superpixel straddling* measure proposed by [8] penalizes windows that have a large number of superpixels straddling their borders. This method uses a faster approach which, instead of working directly with superpixels, operates on the bounding boxes enclosing the superpixels. A binary image is formed consisting only of the borders of windows enclosing all the superpixels in the image. This image is then smoothed using a Gaussian operator. Finally, the score of a candidate window is calculated as the sum of image intensities of this smoothed image along the border of the window being scored, normalized by its perimeter. This approach is much faster to calculate as compared to *superpixel straddling* proposed in [8].

The second objectness measure proposed is very similar to the *edge density* cue introduced in [8] and is termed as *boundary edge distribution* (BE) . Unlike the *edge density* measure which is based on a single edge map calculated using [22], the BE score exploits intensity gradient maps of edges along four directions (horizontal, vertical, 45° and 135°). The objective here is to exploit the closed boundary characteristics of objects. This is achieved by promoting windows having largest weights of edge intensity gradients that are close to the boundary of window and parallel to it. The third and final objectness cue is called *window symmetry* and is based on the simple fact that objects in real-life generally have a symmetric shape. It also makes use of the gradient maps defined in the case of BE.

The features are combined in a structured learning framework as opposed to the Naïve Bayes approach of [8]. A loss function, based on the overlap with ground truth, is optimized. Finally, a Non-Maximum Suppression operation is applied to further reduce the number of proposals by removing spatially redundant ones. Non-maximum suppression, commonly referred to as NMS, aims to reduce redundancy in a set of outputs of an algorithm. It is widely used as a post-processing step in different computer vision applications [4], [11], [22], [26], [27]. The most commonly employed approach is to greedily search for local maxima in the set of outputs and retaining them while suppressing similar results that are possibly redundant.

2.1.3 Binarized Normed Gradients (BING)

Binarized Normed Gradients (BING) [11] is a method which employs fast gradient based features to measure the objectness of windows. The image is scanned for candidate windows having predefined quantized sizes and aspect ratios to populate the initial pool of proposals. Each window is then evaluated based on a learnt objectness measure based on the closed boundary characteristics of objects in images. Finally, an NMS operation is employed to reduce the number of proposals by removing spatially redundant ones.

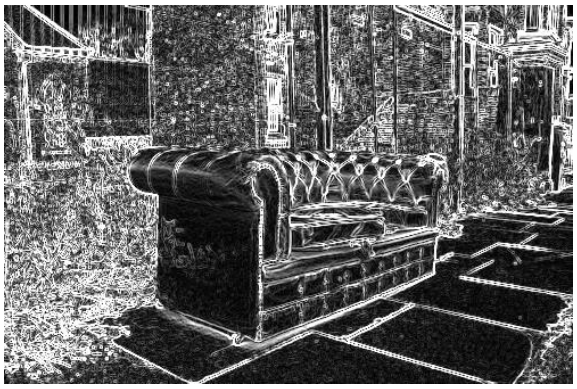

Figure 5. *An example of a gradient image*

Given an input image, a normed gradient (NG) image like the one shown in Figure 5 is constructed by taking gradients in the vertical (g_y) and horizontal (g_x) directions and normalizing them as follows:

$$NG = \min(|g_x + g_y|, 255) \tag{2}$$

The proposed objectness measure revolves around the observation that when shrunk to a very small size (8x8), windows that enclose objects exhibit similar features in the normed gradient image domain. Moreover, these features are easily discriminable from those obtained from windows not enclosing any objects. As observed by the authors, these features, termed as 64-D NG (corresponding to flattened 8x8 windows), are quite robust to changes in color, shape or size of the object because such variations have minimal effect on the directional gradient intensities.

The proposed objectness measure is formulated as follows:

$$s_l = \langle \boldsymbol{w}, \boldsymbol{g_l} \rangle \tag{3}$$

$$l = (i, x, y) \tag{4}$$

In (3), s_l is the output of the filter which takes as input the NG features $\boldsymbol{g_l}$ and coefficients $\boldsymbol{w}$, which are learnt in an SVM framework from ground truth annotations provided in PASCAL VOC 2007 [25] training dataset. (i, x, y) in (4) uniquely identifies a window that has a quantized size i and is located at coordinates (x, y) in the image. As, some window sizes are more probable to contain objects, the final objectness score introduces a bias towards sizes that are most likely to contain objects. It is formulated as given in (5)

$$o_l = v_i . s_l + t_i \tag{5}$$

In (5), v_i and t_i are parameters that incorporate the bias for certain window sizes which are learnt to have a high likelihood of enclosing an object.

BING is a binary approximation of the above proposal generation process which operates at 300fps and yields a good recall in producing windows that provide more than 50% coverage of objects of interest. Figure 6 shows an example image from the PASCAL VOC 2007 dataset and its best proposal obtained using BING. A drawback of this approach is that the windows produced do not tightly enclose the objects of interest. This is evident from Figure 6 which shows a fairly simple image with only a single object yet the best proposal fails to provide an accurate localization.

Figure 6. *Best proposal generated by BING* [11] *(red) compared to ground truth (green)*

2.1.4 Edge Boxes

Edge boxes [18] is another window based method for generating object proposals, which in spite of being unsupervised in its formulation, produces competitive results. This method only utilizes the edges in an image to define the objectness measure. It works on the hypotheses that windows that contain wholly enclosed edge contours are more likely to contain an object. Based on this, the objectness score is designed to measure the strength of contours completely inside the bounding box relative to those which cross its boundaries. Figure 8 shows some examples of the results of the algorithm on a test image taken from PASCAL VOC 2007 dataset.

Figure 7. *Structured forest edge detection* [28]*: Original image (right), edge map (left)*

For an input image, a global edge map is obtained as shown in Figure 7. Given this edge map, the edges that are connected and exhibit a high degree of similarity are grouped

together to form *edge groups*. Given a candidate window, it's objectness score is then calculated using (6).

$$h = \frac{\sum_i w_b(s_i) m_i}{2(b_w + b_h)^\kappa} - \frac{\sum_{p \in b^{in}} m_p}{2(b_w + b_h)^\kappa} \tag{6}$$

In (6), w_b is a real valued indicator of the degree to which an edge group s_i is contained in the group and m_i is the sum of edge magnitudes of all the pixels in the i^{th} edge group. The second term is used to remove the effect of edge groups lying in the center of the box. To this end, a smaller bounding box, b_{in} is defined which is centered on b but has half its width and height. This is based on the observation by [8] that edges in the center of the window hold less objectness information as compared to those on the boundaries.

The initial set of windows is populated using the sliding window paradigm where windows of different sizes and aspect ratios are slid across the entire image. The coarseness of this sliding operation is controlled by a free parameter α. For each of the candidate window, the objectness score is then calculated using (6). Candidates having a score greater than a predefined threshold are further refined by adjusting their position, scale and aspect ratio to maximize the objectness score.

Finally, a non-maximum suppression operation is applied to remove spatially redundant windows. The overlap threshold used for NMS, δ, is kept a free parameter. The exact value of δ is observed to correlate strongly with the performance of the algorithm at IoU (Intersection Over Union) thresholds in the vicinity of δ. An important observation here is that the quality of proposals produced depends heavily on the edge detector used for generating the edge-map. The structured edge forest detector of [29] was observed by the authors to provide best overall performance.

Figure 8. *Best windows (red) proposed by Edge Boxes method* [18] *for the ground truth object (green)*

2.2 Region based Methods

Region-based methods are those methods which produce proposals in the form of multiple segmentations of the image. The objective is the same as in the case of bounding box based methods; to produce a pool of possibly overlapping pixel-accurate regions that have a high probability of covering or enclosing an object.

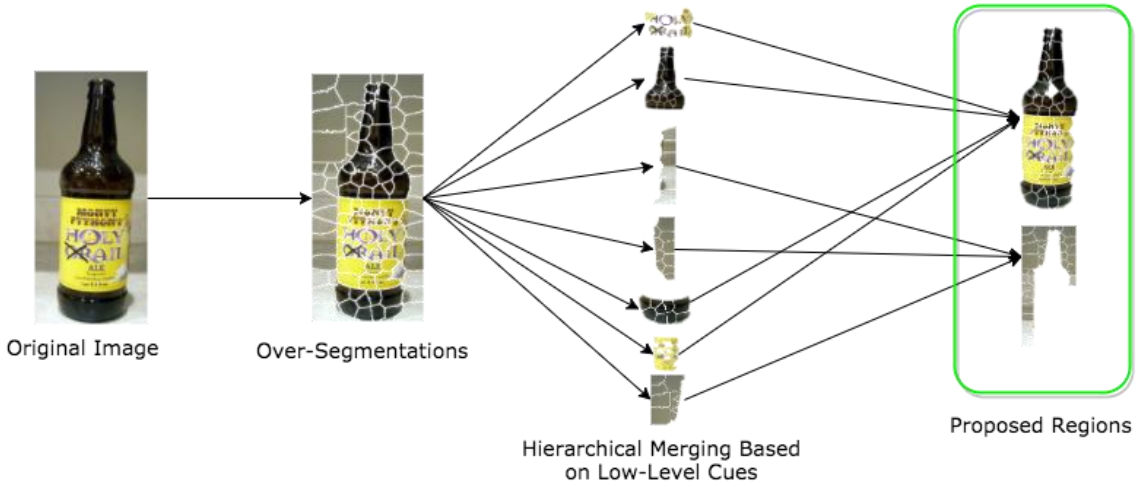

Figure 9. *Grouping based proposal generation methodology*

There exists a wide range of techniques that are used to generate region proposals. Two primary methodologies have emerged in this regard. One approach, as illustrated in Figure 9 is to oversegment the image into small regions and then combining those regions based on various cues based on shape, color or texture [16], [17], [30]. Another approach, as visually demonstrated in Figure 10, is to generate proposals by producing multiple foreground-background segmentations of the image by assuming foreground seeds placed at different locations in the image [10], [13], [14], [31].

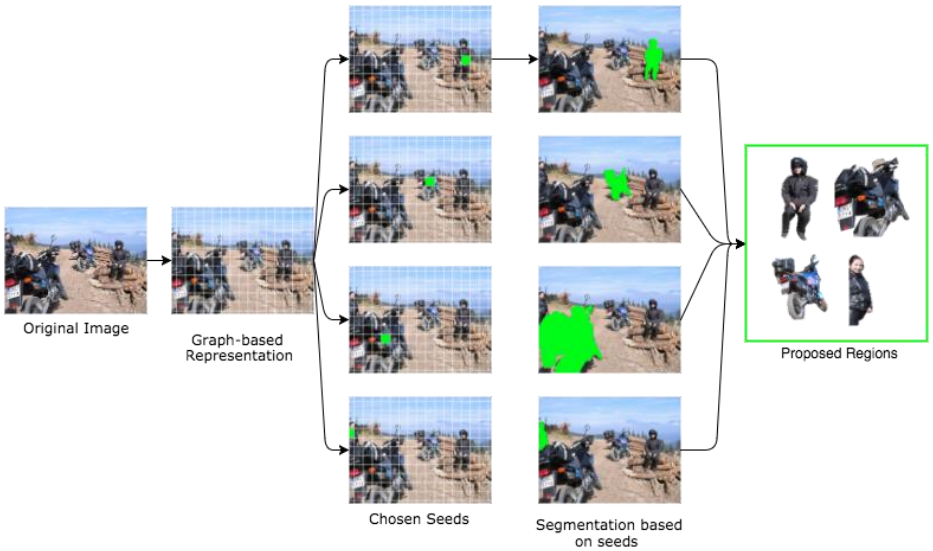

Figure 10. *Seed based proposal generation methodology*

2.2.1 Constrained Parametric Min-Cuts (CPMC)

Constrained Parametric Min-Cuts (CPMC) [10] is one of the first works done in generating a pool of regions providing a high coverage of objects in the image. The problem is formulated in a graph theoretical framework. Regions are generated by solving multiple graph cut problems with different parameters and initializations, each providing a binary segmentation of the image into background and foreground. The uniformly sampled foreground seeds provide the primary source of diversification in the proposals which is further enhanced by applying maximum marginal relevance measures. Furthermore, an objectness scoring function based on low to mid-level cues in the image is also trained and used to rank the proposals.

The pixels of the colored input image are represented as nodes V of a graph $G = (V, E)$ where the edge weights E are representative of the similarity between edges. A set of pixels V_b is provided as a foreground seed and the image boundary is presumed to be belonging to the background seed V_b. Given a binary segmentation of pixels $x_i = \{x_1, x_2 \dots x_k\}$, which is known only for the background and foreground seeds, the objective is to minimize the energy function which takes the form as follows:

$$E^{\lambda}(X) = \sum_{u \in V} D_{\lambda}(x_u) + \sum_{(u,v) \in E} V_{uu}(x_u, x_v) \tag{7}$$

The two summations are over the unary and pairwise potentials respectively and are formulated as below:

$$D_{\lambda}(x_u) = \begin{cases} 0 & if\ x_u = 1\ and\ u \notin V_b \\ \infty & if\ x_u = 1\ and\ u \in V_b \\ \infty & if\ x_u = 0\ and\ u \in V_f \\ f(x_u) + \lambda & if\ x_u = 0\ and\ u \notin V_f \end{cases} \tag{8}$$

$$V_{uv} = \begin{cases} 0 & if\ x_u = x_v \\ g(u, v) & if\ x_u \neq x_v \end{cases} \tag{9}$$

Both the infinity terms correspond to the cases when the one of the seed pixels is mislabeled. The unary weight is zero if the pixel does not belong to the background seed and is labelled as foreground. For the reverse case of labelling a pixel which doesn't belong to the foreground seed as background, the unary cost is not zero. The rationale behind this is to introduce a bias for larger foreground regions. The amount of bias is controlled by the sum of a free parameter λ and a function $f(x_u)$. Multiple real-valued variants of λ are used to explore multiple scales. In addition to that, two variants of the function $f(x_u)$ are also employed; one in which it is kept zero while the other in which its value is proportional to the difference of color histogram distributions of foreground and background seeds.

The pairwise potential term V_{uv} penalizes the cases when two similar pixels are labelled differently. The similarity criterion $g(u, v)$ is based on the contour strengths at the two pixels, calculated using [32]. For a single seed, the problem can be characterized as a parametrized min-cut problem over the graph G with parameter λ, which is then solved by the parametric max-flow solver of [33]. The disconnected components, if present in the foreground segmentation, are split and proposed as separate regions. The set of foreground seeds is obtained by selecting uniform 5x5 rectangular grids over the entire image. The number of seeds, and the λ values explored results in a very large set of initial proposals which is further pruned to maximize diversity while keeping the number of proposals as low as possible.

Figure 11. *Bounding boxes corresponding to the best proposed regions by CPMC* [10] *(red), compared with ground truth (green)*

As a first step towards the reduction of proposals, a *fast rejection* methodology is adopted where all regions having an area less than a predefined limit of 150 pixels are eliminated. Secondly, the segments are ranked based on a simple energy function defined as in [34] and the 2000 segments having the lowest energy are selected. Moreover, the regions produced exhibit a high degree of spatial redundancy. This is primarily because of the fact that as long as the seed remains inside the object, it will produce the same solution. This redundancy is more significant for images having large objects. The effect of this is reduced by clustering the proposals having an IoU overlap equal to or greater than 0.95 and selecting the proposal having the lowest energy value from each of the cluster.

In order to ensure that the regions providing the best coverage of objects are ranked higher, a segment based objectness measure based on low to mid-level cues is trained. The problem is posed as a regression problem over the IoU overlap with the ground truth. A vast set of features is synthesized based on graph partitioning properties, regional properties and Gestalt properties such as inter/intra region continuity. A random forest regressor is employed to regress over the best overlap of a segment with a ground truth object. A benefit of such a setting is that the regressor itself learns which features are more significant towards defining the objectness of the segment being scored. An adverse effect of this ranking strategy is that similar regions, probably belonging to the same object, are placed very close to each other. This is resolved by increasing the diversity of proposals using MMR [35] measures which penalize close placement of similar proposals by incor-

porating a redundancy term when re-ranking proposals. The redundancy measure employed is the IoU overlap with previously ranked proposals. This increases the probability of covering all objects in the image as opposed to only the largest or most appealing one. Figure 11 shows the best proposals obtained using this method for a test image.

RIGOR [13] and Geodesic [14] are two methods that are very similar in formulation to CPMC. RIGOR employs a graph theoretical approach to reuse the graphs for multiple seeds and aims to reduce the computational complexity associated with CPMC. On the other hand, Geodesic method learns the optimal placement of seeds and uses oversegmentations or superpixels as nodes of the graph to obtain a lower number of accurate foreground segmentations.

2.2.2 Multiscale Combinatorial Grouping (MCG)

Multiscale Combinatorial Grouping (MCG) [30] is a relatively recent approach which also produces segment based proposals. It is a unified approach that first generates a pool of segmentations and then groups them in a top-down combinatorial fashion to produce pixel-wise segments. It also offers a choice of several operating points which define a trade-off between the number of proposals generated and the maximum achievable performance in localizing objects. Figure 12 shows the best proposals generated using MCG for a test image.

Figure 12. *Original Image (right), best proposed region of MCG* [30] *shaded green (left)*

The first step of the algorithm is to generate what is termed as an *ultrametric contour map* (UCM). It encodes a family of hierarchical segmentations, arranged from coarse to fine, as a single fused edge map where the strength of contours represent the level of segmentation. An example contour map is shown in Figure 13. Thresholding the contour map at a certain threshold λ_i provides a partitioning that represents the merging of all segmentations below the i^{th} level. The segmentations themselves are generated based on a variety of global and local cues such as brightness, color and edge maps obtained using [29]. The cues are combined linearly using learnt weights. In the multiscale variant of the segmentations, the image size is subsampled and supersampled to different resolutions to maximize diversity of proposals. The UCMs obtained from all the different resolutions are combined using a maximum vote technique to get a global contour map representative of the multiple segmentations.

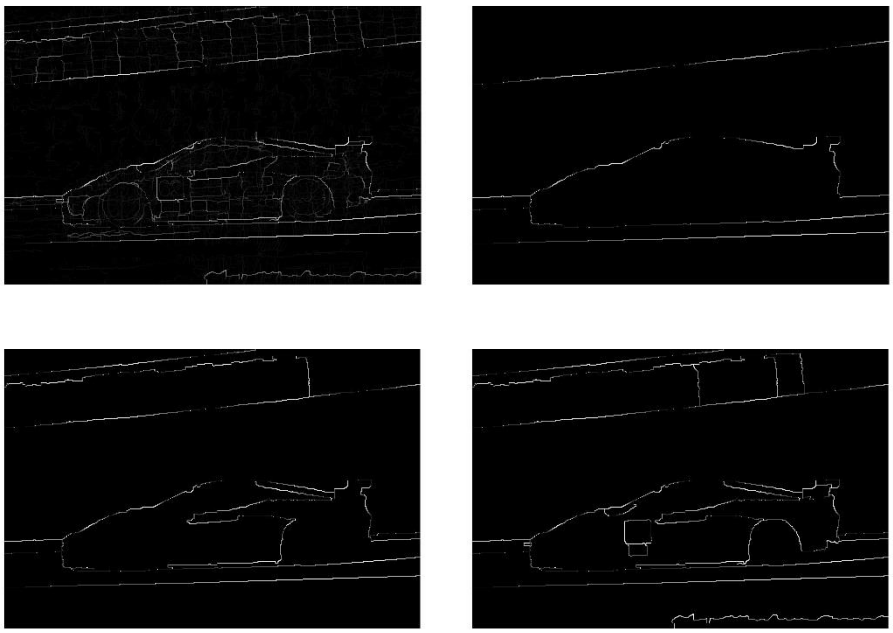

Figure 13. *Ultrametric contour maps: (from left to right, top to bottom) Original UCM, UCMs thresholded at progressively lower levels*

In order to generate object proposals, multiscale segmentation hierarchies are first used to obtain segmented regions, which are then merged in a combinatorial manner. The objective is to propose regions that are more likely to represent a whole object rather than its parts. The depth of the combinatorial sets explored is kept adjustable and provides the trade-off between number of object proposals and the maximum accuracy achieved. This trade-off is modeled as a Pareto Optimization problem where the two contradictory objectives are the number of proposals (lower is better) and the overall accuracy (higher is better).

The generated proposals are also ranked using a random forest regressor, similar to the one in [10]. The task for the regressor is to use low-level features to measure objectness of the regions. The regressor uses various low-level cues based on shape (perimeter, area), compactness and UCM contour strengths. It is trained over PASCAL VOC 2012 [36] dataset. The objectness score obtained from this learnt regressor is used to decrease the number of proposals and diversify them based on a Maximum Marginal Relevance [35] measure.

2.2.3 Selective Search

Selective search [17] is one of the most widely used region-based proposal method. It is essentially a region merging technique where the objective is to combine over-segmentations of an image based on a set of defined similarity metrics. The initial set of regions is obtained from a hierarchical segmentation method, exploring different color spaces and

scales. The aim then, is to combine smaller regions together so as to maximize the likelihood of the combined regions to represent complete objects. To this end, a number of similarity measures are explored to deal with regions of different spatial and textural characteristics. This is achieved by focusing the algorithm for specific design considerations.

Figure 14. *A test image (left) and the best proposed regions by Selective Search* [17] *shaded green (right)*

First, in recognition of the fact that the sizes of objects in an image can vary considerable, a multi-scale approach is found more suitable for the task. Secondly, it is noticed that one single discriminator for defining inter-region similarity is simply not enough for the task of efficient segmentation. Therefore, a set of complimentary features are used together when deciding which regions are to be merged to ensure diversification of proposals. Finally, the process is designed to be computationally efficient so that its use in practical application is better justified.

In order to capture all possible scales, a simple bottom-up grouping strategy is used where a fine over-segmentation of the image is iteratively made coarser until it covers the whole image. Each step of this process essentially represents a different scale of objects. The initial set of regions is obtained using the superpixel segmentation method of [37]. Further diversification is achieved by running [37] with different parameters and on different color spaces. From the initial set of regions, the two most similar regions are then combined to form a new larger region. The inter-region similarity is determined using color (s_{colour}), texture ($s_{texture}$), region size (s_{size}) and geometry (s_{fill}) based cues. The final similarity score between two regions is then represented as follows:

$$s(r_i, r_j) = a1 * s_{color}(r_i, r_j) + a2 * s_{texture}(r_i, r_j) + a3 * s_{size}(r_i, r_j) + a4 * s_{fill}(r_i, r_j) \quad (10)$$

The process is recursively repeated to produce larger and larger regions until the whole image becomes a single region. The aggregated set of regions is finally proposed as the location hypotheses for class-independent objects. The proposed regions are not evaluated and ranked based on any objectness measure. Instead they are ordered in the sequence of generation with an introduced randomness to remove the bias for larger regions. Figure 14 shows the results of Selective Search on an image from PASCAL VOC 2007

data set. The computationally efficiency of the algorithm has seen its use in many state-of-the art object detectors [38], [39],[27].

2.2.4 Rantalankila's Method

Rantalankila's method [16] is another efficient region-based proposal generation method. It builds on the foundations of Selective Search [17], CPMC [10] and Endres [31] to propose a method which incorporates both the local similarities between superpixels like [17] and also diversifies the search in a global context by solving customized graph based problems as in [10], [31] with superpixels represented as nodes.

The method consists of two stages; local and global search. The local search stage is very similar to the approach of Selective Search [17] but different in some critical design parameters. Firstly, the initial oversegmentation image are obtained using both [37] and SLIC [40]. The superpixels produced by the latter are observed to exhibit a higher level of homogeneity as far as their size and shape is concerned and thus acts as a compliment to the more heterogeneous set produced by [37]. Secondly, the similarity measure and the succeeding merging process is different to the approach in [17].

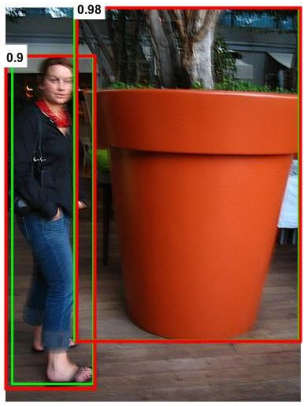

Figure 15. *Bounding boxes corresponding to the best proposals generated using Rantalankila's method* [16] *(red) and the ground truth (green)*

As a first step of the proposal generation process, the image is oversegmented into superpixels using the methods of [40] and [37]. Next, features based on Scale Invariant Feature Transform (SIFT) and RGB color distribution are calculated for each of the superpixel. A pairwise merging of the superpixels then follows where the two most similar superpixels, as per the similarity measure defined earlier, are merged together. Unlike the method adopted in [17] where the merging continues until the whole image becomes a single region and all the intermediate combinations are part of the output, the merging here stops after a certain threshold of similarity is reached and only the final combinations of superpixels are part of the output set. These "refined" superpixels are then combined in an iterative manner to produce the final set of proposals.

The technique described above serves an efficient purpose as far as segmenting locally discriminative regions from the image. However, such a methodology will not be very effective when the object has a closer appearance to the background. In such cases, a global measure is more effective. The second step of this method consists of using all superpixels of the image at once as nodes of a graph. By assigning some nodes as definite foregrounds and backgrounds and minimizing an energy function given by (11), the most optimal binary labelling (foreground/background) for the superpixels is achieved. In the formulation of (11), V and D represent the unary and binary weights for the nodes respectively.

$$E(L) = \sum_{i} D(i, l_i) + \alpha \sum_{i,j \in E} V(i, j, l_i, l_j) \tag{11}$$

The way V and D penalize bad segmentations is similar to the way it's done in [10] as explained in Section 2.2.1. Figure 15 shows the best proposals obtained for an image using Rantalankila's methods.

2.2.5 Randomized Prim's

The Randomized Prim's algorithm [15] is another graph based method that is similar in formulation to CPMC [10]. However, instead of modelling the problem as a parametric min-cut of the graph, Randomized Prim's algorithm aims to *grow* regions from randomized seeds. Prim's algorithm [41] is a greedy algorithm that is used to find spanning trees of graph like structures. The Randomized Prims algorithm uses the same principle to find partial spanning trees of a graph. The main modification to the Prim's algorithm is that instead of greedily sampling for complete spanning trees, Randomized Prim's algorithm aims to maximize the sum of edge weights of partial spanning trees.

Figure 16. *Bounding boxes corresponding to the best proposals generated by Randomized Prim's algorithm* [15] *(red) compared to ground truth (green)*

In order to generate object proposals, the image is oversegmented using [37] into superpixels. In the weighted graph representation, the superpixels act as nodes and they are connected by edge weights which are learnt using logistic regression techniques on the training data. Once the graph structure is formulated, the Randomized Prim algorithm is applied repeatedly to randomly selected starting nodes as follows. For each randomly chosen seed, nodes (superpixels) are repetitively added to the tree by sampling the multinomial distribution of similarities between the nodes in the tree and those in its neighborhood. The growth of tree is terminated when a criterion based on a learnt stopping function is reached. Finally, the tightest bounding box enclosing the superpixels in the tree is proposed as an object proposal. Repeating this process over multiple random seeds aims to produce multiple diverse object proposals which have a high probability of covering all objects of interest.

For a pair of neighboring superpixels n and m, the edge weight, $\rho_{n,m}$, aims to model the probability that they belong to the same object. It is formulated as follows:

$$\rho_{n,m} = \sigma(\boldsymbol{w}^T \phi_{nm} + \boldsymbol{b}) \tag{12}$$

In (12), $\boldsymbol{w}$ represents the weights that are learnt on the training data, $\phi_{nm}$is the feature vector and $\boldsymbol{b}$ is the bias term. σ represents the sigmoid function. The feature vector is comprised of three similarity measures. First is the color similarity f_c which is based on normalized histograms in the Lab color space. Second is termed as *common border ratio* f_b which is the maximum ratio of the length of borders shared by the two superpixels normalized with their perimeters. For superpixel segmentation algorithms like [37] which are more biased towards color consistency in superpixels as compared to their form-factor, the common border ratio is a strong measure of their probability of belonging to the same object. The last feature f_s is based on size and favors the grouping of two small superpixels over larger ones. In the training phase, superpixels are annotated as belonging to an object if more than 60% of their area is inside an object. The final values of weights and biases are calculated by maximizing a log likelihood estimate.

$$\{w^*, b^*\} = argmax \sum_i y_i \ln p_i + (1 - y_i)\ln(1 - p_i) \tag{13}$$

The termination function takes the form of (13). It consists of addition of two terms. The first one, $(1 - p_i)$ is the probability that the last sampled edge is connecting two superpixels that do not belong to the same object. The second term $\alpha(T_k)$ is the fraction of objects in the training data whose size is smaller than T_k. This aims to prevent the size of proposal from growing larger than the annotated objects in the ground truth training data. Figure 16 shows the result of this method on a test image containing multiple overlapping objects of interest.

2.2.6 Endres's Method

The method proposed by Endres et al in [31] is also based on multiple foreground/background segmentations to generate object proposals. The seeds used for the segmentations are not random as in the case of [10], [15]. Instead, they are sampled from a hierarchical segmentation that is obtained using the occlusion boundary maps inferred by [42]. The initial pool of regions obtained is then refined using an extensive ranking strategy that aims to simultaneously maximize the coverage of object and minimize the redundancy between proposals to increase diversity.

The occlusion boundary method of [42] provides a series of coarser segmentations of the image. Each of these segmentations are used to construct a pixel-wise boundary map and a foreground probability map of the entire image. These segmentations are then averaged to get an average boundary and figure/ground probability maps. Regions from the average map are merged based on the strength of their boundaries to produce a hierarchical segmentation of the image. Seeds for generating proposals are then chosen from this segmentation based on their sizes and the strength of boundaries. Larger regions are preferred over smaller ones and regions with weak strength are ignored on the presumption that they belong to the interior of an object.

The generation method consists of finding regions that are similar to a chosen seed and therefore, have a high probability of belonging to the same object as the seed. For this purpose, a conditional random field (CRF) is employed that infers the binary labeling $l_i \in \{0,1\}$ (foreground, background) over all superpixels. This is formulized as in (14)

$$P(\boldsymbol{l} \mid \boldsymbol{X}, \boldsymbol{S}, \gamma, \beta) \propto \exp\Big(\sum_i f(l_i; \boldsymbol{S}, \boldsymbol{X}, \gamma) + \beta \sum_i g\big(l_i, l_j; \boldsymbol{X}\big)\Big) \tag{14}$$

where $f(l_i; S, X, \gamma)$ and $g\big(l_i, l_j; X\big)$ are local and pairwise potentials defined as superpixel affinity and edge cost respectively. Here, $\boldsymbol{X}$ is a set of all image features and $\boldsymbol{S}$ is the set of superpixels belonging to the seed. The foreground bias γ and affinity/edge cost tradeoff β are free parameters which are varied in predefined intervals for each seed.

The affinity is modelled by first learning the probabilities of the set of regions $\boldsymbol{R}$ to lie on the same object as the seed $\boldsymbol{S}$. The features used for this purpose are based on a variety of local cues consisting of color and textural histograms, relative boundary strengths and geometrical layout agreements. The positive examples are synthesized using a pair of regions which lie on the same object and the negative examples are generated using the pairs which do not. The region-wise probabilities are transferred to superpixels by simple averaging normalized by a homogeneity measure for the region. Finally, the superpixel-wise affinities are used to formulate the affinity cost $f(l_i; \boldsymbol{S}, \boldsymbol{X}, \gamma)$. For the pairwise edge cost, the boundary maps are employed to penalize the cases where two adjacent superpixels with a relatively low boundary strength are classified as having different labels.

For each seed and the parameter combination, the CRF produces an exact inference using the graph cut technique of [43]. By separating the regions having disconnected components and removing those having high spatial redundancy, a final pool of diversified region proposals is obtained.

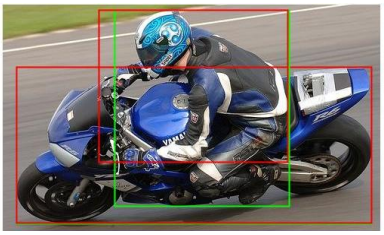

Figure 17. *Bounding boxes for best proposals obtained using Endres's method* [12] *(red) and the ground truth (green)*

The second step of the process consists of ranking the proposals. The objective here is to jointly optimize the class-independent objectness of proposals and their diversity. This is done by greedily maximizing the scoring function of the form shown in. It is a blend of two functions $\psi(\boldsymbol{X})$ and $\phi(r)$ weighted by a monotonically decreasing function $\alpha(r)$. The former is a function of **X** (the set of all proposals) and represents the appearance features while the latter represents inter-region overlap and aims to increase diversity by discouraging spatial redundancy. The appearance features are learnt based on color and textural distances of ground truth annotation of objects and backgrounds. Finally, the scoring function, $\boldsymbol{S}(x, r; w)$ is maximized by greedily adding the proposal with the maximum marginal gain. The optimal ranking r_o is learnt in a structured learning framework. Figure 17 shows the most accurate proposals obtained using the method and their overlap with the ground truth, shown in green.

2.3 Qualitative Evaluation of Methods

The previous section chronicles the state-of-the-art object proposal generation methods. Bounding box based methods generally sample a large number of rectangular windows from the image and rank them based on an objectness measure. On the other hand, all the region-based methods propose pixel-wise accurate segments. The use of bounding boxes to localize regions significantly decreases the computational overhead associated with generation, evaluation and non-maximum suppression of proposals. That's the main reason behind the fast performance of BING [11] and Edge Boxes [18], as shown in Figure 18. However, in spite of their popularity in various detection and classification frameworks, it can be intuitively argued that using bounding boxes to localize objects is perhaps not optimal. This results from the fact that objects in real life do not necessarily have a perfect rectangular shape and enclosing a bounding box around a non-rectangular object will inevitably allow some pixels that are not part of the object, even when the bounding box is the tightest possible.

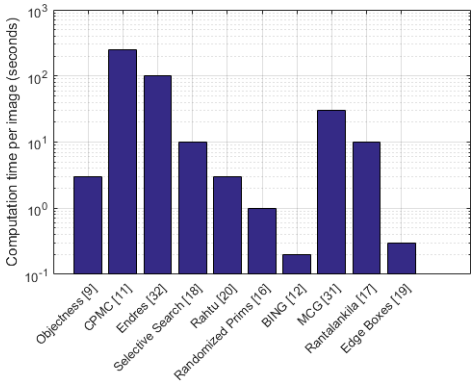

Figure 18. *Comparison of computational times for proposal generation methods*

In case of objects having symmetric shapes and box-like form factor, bounding boxes work quite well for all practical applications. However, in the case of more complex structures, a bounding box based localization fails to capture the intricacies of the object. This is further elaborated with the help of the example in Figure 19.

The top row in Figure 19 shows an image where the object of interest has an approximately rectangular and compact shape. In this case, bounding box provides a satisfactory localization. The bottom row shows an image with a complex structured object (snake) along with both bounding-box and region based localization. The box shown is the tightest fit possible for the object and further shrinkage of the box from either sides will result in some part of the object being outside the boundaries of the box. We can clearly see that more than half of the pixels enclosed by the bounding box do not belong to the object. This coarseness in the bounding-box based localization is undesirable in many applications such as robotics which require accurate details about the spatial extent of an object to perform physical tasks like grabbing. Hence, in an ideal class-agnostic framework, where the objects of interest can be of all shapes and sizes, segment based proposals provide much better localizations.

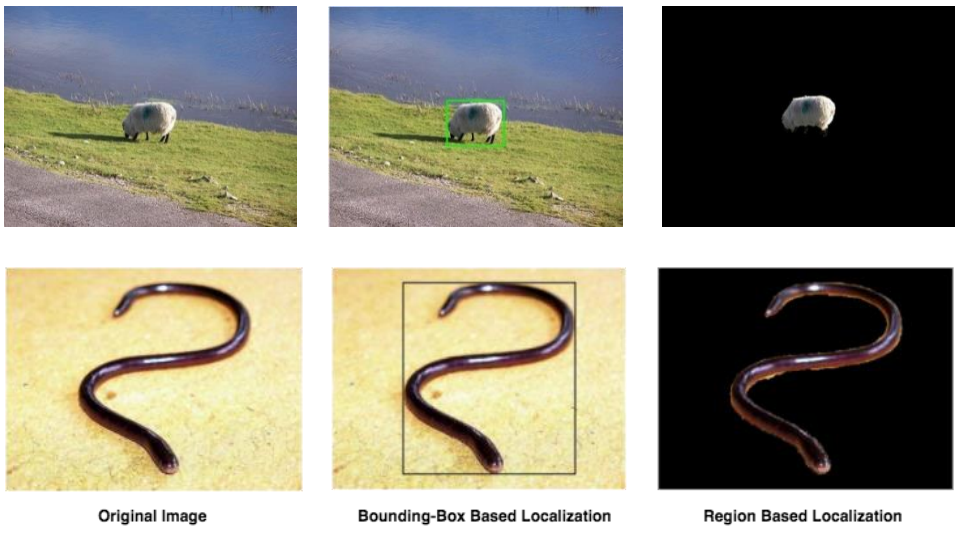

Figure 19. *Comparison of bounding-box (2nd column) and region based localization (3rd column) for objects in two example images (rows)*

Diversification of proposals is another pertinent issue when dealing with category independent proposal generation. A diversified set of proposals ensures coverage of all objects present in the image, irrespective of their class or prominence. Bounding-box based methods achieve this by exploring windows spread over the entire image and exploring a variety of aspect ratios and sizes of windows. However, in almost all the methods, the objectness measure is learnt over some training data. Therefore, there is a strong probability that the objectness scores will overfit the classes present in the training data or those visually similar to them.

Segment based proposal methods take different routes to ensure diversification of proposals. Methods like MCG [30] and Selective Search [17] which essentially merge fine over-segmentations of the image achieve this by adjusting the coarseness of initial over-segmentations, exploring multiple color spaces and employing a hierarchical merging process. However, as noticed by [16], such a *local search* is not suitable for images with large non homogeneous objects. In such cases, an operation that considers all parts of the image in a global context can provide better results for different types of objects. In methods employing such a global operation [10], [12]–[16], diversity is achieved by exploring different schemes of initializing the foreground seeds. The authenticity of these seeds is the single deciding factor of the quality of proposal obtained. A misplaced seed can ruin the subsequent segmentation and produce proposals that may not belong to objects at all. To compensate for this, most of the methods also employ a supervised framework, either to learn the efficient placement of seeds or to learn an explicit objectness function.

From the above discussion, we can summarize some of the problems faced with current generation strategies as follows. Bounding box based methods produce proposals that provide a coarse localization of the object but their generation is very fast and post-processing operations like non-maximum suppression are computationally economical. Segment based proposals methods aim to produce pixel-wise segments which allows them to effectively localize objects with complex geometry. However, to ensure that objects of all classes and sizes are present in the output set, dependence on learnt or supervised schemes becomes imminent.

Keeping these points in view, it can be argued that an ideal proposal generation scheme should be able to localize the object boundaries well, be computationally efficient and must be devoid of any learnt or supervised parameters, for learning inevitably makes the proposals biased towards objects of certain classes. To this end, we look for a fast segmentation method that is unsupervised and does not make any category-dependent assumptions about the foreground.

2.4 Quantum Cut

2.4.1 Background

Graphs are structures that model pairwise similarities between entities. They are a collection of points called *vertices* or *nodes* and pair-wise connections between them, referred to as *edges*. In a weighted graph, an edge joining a node i to another node j, also has a weight associated with it represented by w_{ij}. An undirected graph is a weighted graph in which the edges are bi-directional i.e. ($w_{ij} = w_{ji}$). Any set of points $x = \{x_1, x_2, x_3 \dots x_n\}$ in an n-dimensional space with predefined pairwise similarities between them can be represented as a graph where the nodes of the graph represent the data points and the edge weights are proportional to the similarity between the two points.

Graph-based methods have shown considerable success in many areas in computer vision, especially clustering. Clustering is a useful data-analysis technique where the objective is to group a given set of data points into different clusters. The data points grouped together in the same cluster are desired to be as similar as possible while those in different clusters are desired to be as dissimilar as possible. In a graph-based representation, clustering corresponds to obtaining a partition of the graph. Consider a specific case where the data points need to be divided into two clusters. This is analogous to obtaining a partitioning of the graph into two-subgraphs. One way of quantifying such a partition is by a measure called *cut* which is defined as the summation of edges that are disconnected as a result of the partition. It is formulated as follows:

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij} \tag{15}$$

In (15), A and B are the two disjoint clusters of nodes that result from the partition. A simple method to obtaining a partition that ensures inter-cluster disparity is to minimize the cost given in (15). This is known as the *min-cut* approach [44]. However, such a partitioning is biased towards producing abnormally small partitions e.g. separating a single vertex from the graph. Keeping this in mind, the cut cost is desired to simultaneously maximize inter-cluster disparity and intra-cluster similarity. Ratio Cut [34] and Normalized Cut [45] propose two different solutions to this problem.

Ratio cut aims to achieve this by ensuring that none of the partition is abnormally small. It proposes to minimize the cost function given in (16).

$$RatioCut(A, B) = \frac{cut(A, B)}{|A|} + \frac{cut(A, B)}{|B|} \tag{16}$$

The size of a partition is defined as the number of nodes present in it and represented by the $|.|$ operator in (16). The cut cost will be increased when either A or B are very small.

The cost function proposed by Normalized Cut ensures that nodes contained in a cluster have strong connections between them. The cost function is defined as:

$$NormalizedCut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)} \quad (17)$$

In (17), the $vol(.)$ operator represents the sum of weights of edges contained in a partition.

Obtaining an exact solution for minimizing the cost functions given in (16) and (17) is an NP hard problem. However, an approximate solution can be obtained using spectral methods that solve the minimization problem by eigen-decomposition of matrices describing the graph.

Graph-based clustering described above have been used to solve image segmentation problems. Any image can be represented as a graph where the nodes of the graph represent the pixels/superpixels of the image and the edge weights are proportional to the similarity between two pixels/superpixels based on e.g. their color. An efficient partitioning of this graph will consequently produce useful segmentations of the image.

2.4.2 Formulation of Quantum Cut

Both the methods of spectral graph clustering described in the previous section, Ratio Cut and Normalized Cut, can be generalized for obtaining as many partitions of the graph as possible. The cost functions are designed to ensure that each of the resulting clusters have a uniform distribution of nodes (Ratio Cut) or high intra-cluster affinities (Normalized Cut). Although both of these methods provide accurate partitions, neither of them are specialized for the specific case of separating foreground from background in images. Quantum cut [46], abbreviated as QCUT, is a recently proposed graph-based clustering technique that presents a solution to this problem.

In QCUT, as in the previously discussed methods, the image is transformed into a graph based representation where pixels/superpixels act as nodes of the graph which are connected by edge weights proportional to their similarity in the Lab color space. In addition, a synthetic node “BKG” is added to the graph which represents the background. Instead of partitioning the graph into two similar clusters as in the case of [34] and [45], the goal here is to extract the foreground segment A which ideally has a large area and distinct appearance from its surroundings.

Both these requirements are met by minimizing the cost function given in (18):

$$\frac{cut(A,\bar{A})}{|A|} \quad (18)$$

The denominator term in (18) ensures that the foreground segment is large enough while the $cut(A, \bar{A})$ term ensures distinctness from the background. Note that this formulation is different from those of Ratio Cut and Normalized Cut in the sense that it only aims to make the foreground segment large and distinct while not imposing any restrictions on the background. This makes it more suitable for foreground extraction in images.

The cut term $cut(A, \bar{A})$ takes into account both the inter-node connections (pairwise potential) and the connection to the synthetic background node BKG (unary potential). It is formulated as follows:

$$cut(A, \bar{A}) = cut_u(A, \bar{A}) + cut_B(A, \bar{A}) \tag{19}$$

$$cut_B(A, \bar{A}) = \sum_{i,j} w_{ij} \left(y_i(1 - y_j)\right) \tag{20}$$

$$cut_u(A, \bar{A}) = \sum_{i} V(i) y_i \tag{21}$$

$cut_B(A, \bar{A})$ penalizes for the cases when two similar pixels are assigned different labels while $cut_u(A, \bar{A})$ penalizes for the case when a node having strong connection with the BKG node is assigned to the foreground. Here $\boldsymbol{y}$ is a binary indicator label which takes value 0 for nodes labelled as background and 1 for those belonging to foreground. The unary potential term $V(i)$ encodes the background prior information which corresponds to the strength of connection with the background node BKG. For nodes that are known beforehand to belong to background, a high value of V is set and for all other nodes $V(i)$ is set to zero. For the application of salient object segmentation, the boundary prior is employed which works on the assumption that pixels on the boundary of the image belong to the background. Apart from that, no strong priors about the foreground or the background are used.

As observed in [46], the minimization of the cost function given in (18) is equivalent to solving the problem given in (22)

$$argmin_{\boldsymbol{z}} \frac{\boldsymbol{z_T}(\boldsymbol{H_m})\boldsymbol{z}}{\boldsymbol{z_T z}} \tag{22}$$

In (22), $\boldsymbol{z}$ is a vector such that $\boldsymbol{y} = \boldsymbol{z} \odot \boldsymbol{z}$ where $\odot$ represents the element-wise multiplication or Hadamard product. $\boldsymbol{H_m}$ is given as:

$$\boldsymbol{H_m}(\boldsymbol{i}, \boldsymbol{j}) = \begin{cases} V(i) + \sum_{k \in N_i} w_{ik} & if\ i = j \\ -w_{ij} & if\ j \in N_i \\ 0 & otherwise \end{cases} \tag{23}$$

Where N_i is the set of neighbours of the node i. Obtaining an exact value of z_i that solves (22) makes the problem N.P. hard. However, if we relax the constraints by allowing z_i to take real-values, a solution can be obtained using spectral methods. The vector $\boldsymbol{z}$ minimizing the above criterion will then correspond to the eigenvector corresponding to minimum eigenvalue of $\boldsymbol{H_m}$ i.e.

$$\boldsymbol{H_m z^*} = E_m \boldsymbol{z^*} \tag{24}$$

In (24), E_m is the smallest eigenvalue of $\boldsymbol{H_m}$ and $\boldsymbol{z}^*$ corresponds to the globally optimum solution. As is clear from (23), non-zero values for the unary cut cost cut_u will ensure that $\boldsymbol{H_m}$ is a Hermitian and positive definite matrix which means that all of its eigenvalues will be greater than zero. The globally optimum solution, $\boldsymbol{z}^*$is then the eigenvector corresponding to the smallest eigenvalue. Reversing the notation change from $\boldsymbol{y}$ to $\boldsymbol{z}$, the optimal labelling vector $\boldsymbol{y}^*$ is now given as

$$\boldsymbol{y^*} = \boldsymbol{z^*} \odot \boldsymbol{z^*} \tag{25}$$

As explained in detail in [47], apart from the eigenvector corresponding to the smallest eigenvalue, other eigenvectors corresponding to progressively higher eigenvalues can also produce ”useful” clustering results. Furthermore, as observed in [46], the globally optimum labelling vector is similar to the solution of discrete time-independent Schrodinger’s equation and is equivalent to the probability density function of a particle at ground state. Both these observations are exploited in Section 3.2 to extend the search for objects.

2.4.3 Performance in Visual Saliency Estimation

QCUT has been applied in a variety of image segmentation problems, most notably in visual saliency estimation [48] and salient object extraction [46][49]. In a comparison conducted over around 41 different methods and 7 datasets, EQCUT [46], a multi-resolution extension of QCUT, has the best performance among all unsupervised methods [50]. Moreover, EQCUT gives state-of-the art results in saliency estimation over a set 6 different datasets as presented in [48] and [49]. Also, in [51], a multispectral variant of QCUT is employed to generate and rank salient segments in an image. The results in that study also put QCUT amongst the state-of-the-art methods. Figure 20 shows multiple images and their saliency maps obtained using EQCUT.

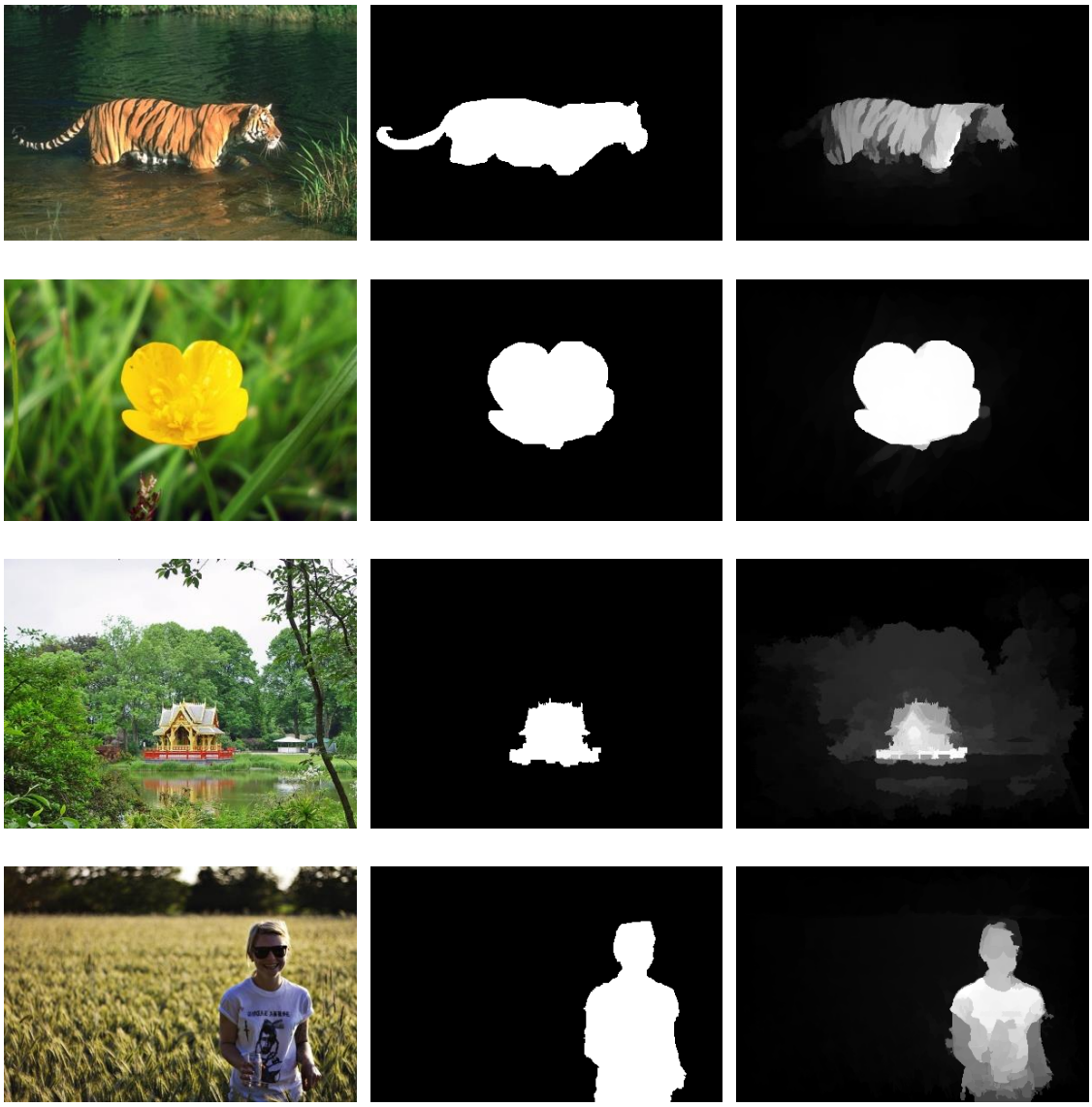

Figure 20. *Saliency Estimation results of EQCUT: (from left to right) Original Image, Ground Truth, EQCUT output*

3. PROPOSED METHOD

3.1 Proposal Generation

In the previous section, the successful application of QCUT in visual saliency estimation and salient object extraction was presented. The objective in saliency estimation problems is to highlight the most appealing part of the image which is a natural center of attention for the human visual system. This differs from the aims of object proposal generation in certain key aspects. Firstly, the task in object proposal generation is to separately localize each instance of an object in the image whereas in saliency estimation, the objective is to provide a pixel-wise probability map of the image for the likelihood of belonging to the salient part. Secondly, all objects in an image are not necessarily salient. An image can contain many such objects which are not visually appealing or not at the center of attention. In the ensuing passages, a visual example is presented for each of these cases to further elaborate this.

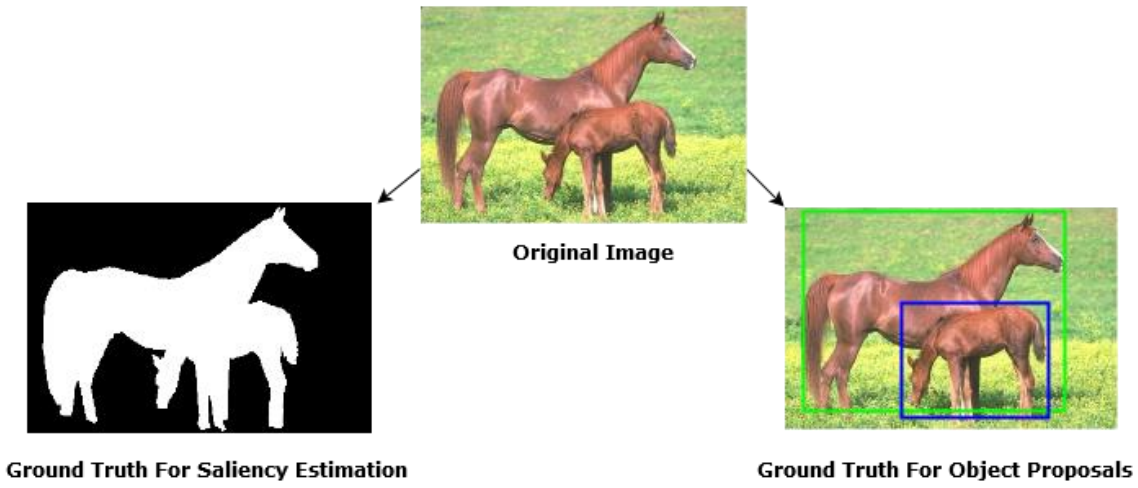

Figure 21. *Differing objectives of Saliency Estimation and Object Proposals*

Figure 21 shows a natural image of two horses standing in a green field. The left image in the bottom row shows the ground truth for saliency estimation. As is clear from this image, the objective is not to localize each of the horses but rather to provide a binary labelling of the pixels based on whether they belong to a salient part of the image or not. On the other hand, the image on the right illustrates the required output from an objective proposal generation system. The aim here is to localize the two horses separately and provide an accurate pool of proposals that provide sufficient coverage for each of the two horses. To further elaborate the point that not all objects in the image are salient, Figure 22 shows an image from the PASCAL VOC 2007 [25] dataset and its corresponding ground truth for object proposal generation. We can see that the person in the background is not at the center of visual attention and hence not a salient part of the image. However,

it is an annotated object and hence an accurate object proposal generation method should be able to provide localizations for it as shown in the last column of Figure 22. Another thing to note here is that a salient region is almost always a part of the output required for object proposal generation. The saliency of a region therefore, acts as an important objectness cue and it has been used for this purpose by [8] when defining a class-agnostic objectness function.

Figure 22. *Not all objects are salient: (from left to right) Original Image, Saliency Map, Object Proposal ground truth (green)*

In light of the above discussion, we modify the original EQCUT method as proposed in [48], [46] in order to make it more suitable for the application of generating object proposals. We relax the background prior by decreasing the boundary potential $V(i)$ which results in larger foreground regions and increases the likelihood for covering objects near the boundary of the image. The effect of this modification can be seen in Figure 23. The second column in the figure is the EQCUT output for the image shown in the first column. As seen in the image, the cows on the left side of the image are not highlighted. However, after decreasing the boundary potential, the comparatively less salient parts are highlighted as well.

Figure 23. *Modifications on EQCUT: (from left to right) Original Image, EQCUT output, EQCUT output with decreased boundary potential*

The goal now is to extract object proposals from the saliency maps like the one shown in the last column of Figure 23.

As a benchmark, we define three characteristics of a good proposal extraction technique:

- Computationally efficient; given a saliency map, the generation of object proposals should be as fast as possible. A fast method ensures the usage of multiple saliency maps without making the whole process prohibitively expensive.

- The proposals obtained must be able to separate out and localize each instance of an object. In the case of the example presented in Figure 24, each the two horses must be localized.
- Regions having high saliency must be represented more often in the object proposals as there's a high certainty for them belonging to an object. On the contrary, lower saliency regions must not be represented as often. However, it is not wise to completely reject them, because of the fact that they might yet contain objects of interest.

To fulfill the above criteria, two different approaches are applied on the saliency maps and presented below. A quantitative comparison of these two approaches is made in Section 4.2.1.

3.1.1 Adaptive Thresholding Based Proposal Generation

Given a two-dimensional saliency map S of the image, a binary image B is generated as follows:

$$B(i,j) = \begin{cases} 1 \; if \; S(i,j) \geq \tau \\ 0 \; if \; S(i,j) < \tau \end{cases} \tag{26}$$

The connected component in B having the largest area is then proposed as an object proposal. The threshold value τ is not fixed and is calculated anew for each saliency map using the Otsu's method [52]. The algorithm clusters the gray-scale pixel values in the image, based on their histogram, such that the inter-cluster variance is maximized. It is used in a variety of image processing applications which involve automatic conversion of gray-scale images to binary format.

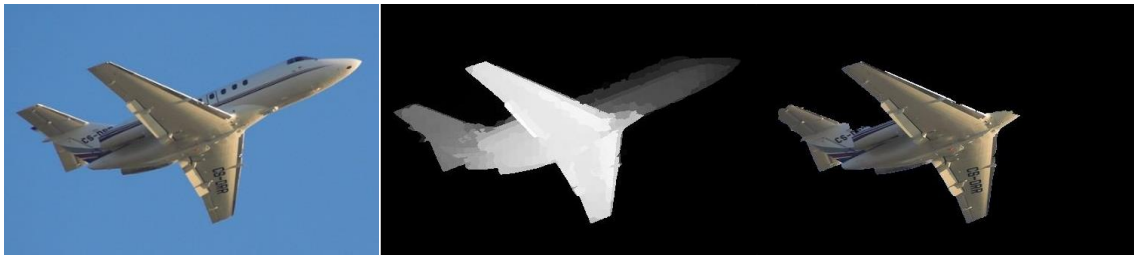

Figure 24. *Proposal extraction using Adaptive Thresholding on Saliency Maps: (from left to right) Original Image, Saliency Map, Proposal extracted through adaptive thresholding*

This method of extracting proposals from saliency maps is very fast as it essentially involves a single thresholding operation followed by connected component labelling. It takes QCUT 2.5 seconds on average to generate 50 eigenvectors. On top of that, adaptive thresholding based proposal extraction takes less than 0.01 seconds per saliency map. One drawback of this method is that it only works well in the cases when the objects of interest

are homogenous w.r.t their saliency and not conjoined with each other. Consider the example shown in Figure 24. The object is composed of several regions, each of which has a slightly different average saliency. An adaptive thresholding operation cuts some part of the object and the resulting proposal as a result lacks accuracy.

3.1.2 Multi-level Thresholding Based Proposal Generation

Realizing the drawbacks of using a single threshold value, a natural extension of the previous approach is to employ a multi-level thresholding. Given a saliency map $S \in [0,1]$, it is first quantized into 256 grayscale levels. This quantized image is then thresholded at all possible grayscale values $\tau_i \in \{0,1,2,\dots,256\}$. Hence, for each threshold level τ_i we get a binary image B_i as follows:

$$B_i(i,j) = \begin{cases} 1 \; if \; S(i,j) \geq \tau_i \\ 0 \; if \; S(i,j) < \; \tau_i \end{cases} \tag{27}$$

Each of the binary image B_i is then subjected to a connected component labelling. All the connected components from each of the binary images is added to the pool of proposals. This is somewhat analogous to performing a bottom-up grouping of superpixels but instead of exploring a vast combinatorial space like [17] and [30], we greedily make the regions finer by maximizing their average saliency. This approach is more expensive as compared to that of the previous section as it involves multiple thresholding and connected-component labelling operations per saliency map instead of one. The computation time per saliency map was observed to be 0.5 seconds on average.

The efficiency of this method in extracting proposals from saliency maps is demonstrated further in Figure 25. Thresholding the saliency map at progressively smaller values increases the chances of producing more complete and accurate object proposals.

Figure 25. *Multi-level thresholding based proposal extraction: (from left to right) Original Image, Saliency Map, Proposals generated through multilevel thresholding*

This method also satisfies the last criterion defined in Section 3.1 i.e. the high saliency regions are represented more often because of the nature of thresholding. Moreover, the regions with low saliency values will not be entirely ignored. This is demonstrated in Figure 25 which shows that some parts of the object of interest can have low saliency values. A global thresholding operation will therefore produce proposals that do not completely enclose the object and occlude some part of it. Multi-level thresholding solves this

particular problem. Drawbacks of this method are the increased computational complexity and the generation of numerous spatially redundant proposals. The effects of the latter problem can be resolved by employing non-maximum suppression.

3.2 Extending the search for objects

As presented in Section 2.4.2, an important property of QCUT which differentiates it from traditional spectral clustering or saliency detection techniques, is its link with quantum mechanical principles. We revisit the formulations presented in Section 2.4.2 and elucidate their usage in increasing the diversity of proposals using this link.

3.2.1 Multiple Eigenstates

As discussed in Section 2.4.2, the optimum labelling vector for the task of saliency estimation was observed to correspond to the probability density function of a particle's position in space at its minimum energy or ground state. This is represented by the eigenvector corresponding to the minimum eigenvalue. However, the eigenvectors belonging to other subsequently higher eigenvalues are also solutions to the Schrodinger's equation and hence valid states of the quantum system. Therefore, while still restricting ourselves to lower energy states, we can exploit the corresponding saliency maps to expand our search of objects beyond the salient ones. It must be noted here that the term "saliency map" is being used only for notational consistency. It shouldn't be confused by the globally optimum saliency map, which has been established to correspond to the ground-state eigenfunction. The saliency maps obtained from other states can be thought of as locally optimum solutions to the saliency problem.

Figure 26. *Multiple eigenstates: (left-to-right) Original Image, Saliency maps corresponding to progressively higher eigenvalues*

Figure 26 shows the saliency maps obtained from the eigenvectors corresponding to progressively larger eigenstates. As shown in the images, the higher energy states also correspond to useful segmentations which can be utilized to obtain better coverage of objects.

Such a multi-spectral approach has also been employed earlier by [51] in generating segments for salient objects.

3.2.2 Quantum Superposition

The principle of superposition states that an eigenstate $\breve{\psi}$ can be expanded as a linear combination of normalized eigenstates as shown in (28) where $\psi_1, \psi_2, \psi_3, \dots, \psi_n$ constitute the basis of the space occupied by $\breve{\psi}$.

$$\widetilde{\psi_k} = c_1^k \psi_1 + c_2^k \psi_2 + \cdots + c_n^k \psi_n \tag{28}$$

In (28), $c_1^k, c_2^k, \dots, c_n^k$ are arbitrary coefficients satisfying the normalization criterion. By virtue of linearity, if $\psi_1, \psi_2, \psi_3, \dots, \psi_n$ are solutions to the Schrodinger's equation, $\breve{\psi}$ is also a valid solution.

The formulation of QCUT allows us to utilize this principle in our framework. As $\boldsymbol{H_m}$ is symmetric, all of its eigenvectors are orthogonal and thus form an orthonormal basis. Therefore, a superposition of these eigenvectors as formulated in (28) can represent any point in space. However, as we are interested only in the salient parts of the image, we only superpose the lower energy eigenstates. Furthermore, to increase the likelihood of generating valid foreground segments, we learn the superposition coefficients as described next.

3.2.2.1 Superposition Coefficients

Given a binary mask of a ground truth annotated object in an image, it is mapped into superpixels by taking mean intensities of pixels of the mask lying on a superpixel. We then calculate the coefficient for the i^{th} smallest eigenvector as in (29)

$$c_i = b_i \cdot \phi_i \tag{29}$$

ϕ_i in (29) represents the i^{th} smallest eigenvector, b_i is the ground truth mask mapped into superpixels and · represents the dot-multiplication operation. We confine the superposition to 10 smallest eigenstates for all images. The process is made over all ground truth annotations to get a set of vectors $\boldsymbol{c} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{c_3} \dots \boldsymbol{c_n}\}$ where n is the total number of annotated objects in the dataset and each vector $\boldsymbol{c_j}$ consists of 10 scalar coefficients corresponding to a unique superposition. It must be noted here that we do not learn any information about the visual appearance and characteristics of the ground truth objects, but rather use the masks to obtain useable combinations of basis vectors that produce valid foreground segmentations. This is a very *shallow* form of learning and doesn't infer any knowledge about the class of the object.

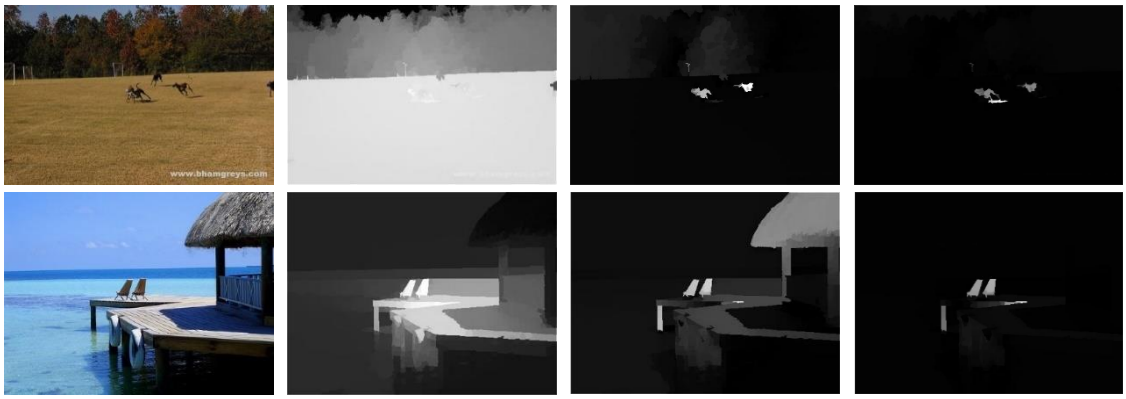

Figure 27. *Superposition of eigenstates: (from left to right) the Original Image, Saliency Map corresponding to the lowest eigenstate and two unique superpositions*

We can see from Figure 27 that superpositions provide saliency maps that enable us to explore different regions of the image in the search for objects. In the first row, the two dogs are very clearly highlighted in the saliency maps obtained through superpositions. Also, in the second row, the three superpositions highlight the hut and the chairs separately which ensures that both the objects are successfully extracted. This is the first application of quantum superposition principle in the realm of image processing applications.

3.3 Ranking of Proposals

As a result of processing multiple eigenstates and their superpositions, an initial pool of proposals is obtained. However, these proposals exhibit a high degree of spatial redundancy. Moreover, many proposals compromise trivial segmentations which is a side-effect of proposing disconnected regions separately. Therefore, it becomes a necessity to decrease the number of proposals to a manageable number. A greedy non-maximum suppression operation can be used to decrease the number of proposals. However, this comes at the cost of a significant loss in recall. As a workaround, we propose a two-pronged strategy which consists of ranking proposals based on a fast unsupervised objectness measure followed by a greedy non-maximum suppression. This enables us to achieve sufficient reduction in the number of candidates while keeping the loss of recall to a minimum. The subsequent sections provide details about each of these two operations.

3.3.1 Unsupervised Ranking of Proposals

We define a fast and unsupervised scoring function of proposals that is parameter free and provides competitive results with the parameter-dependent and learnt classifiers of most of the methods described in Section 2. We use a variety of local cues based on saliency, edge density and compactness as described below:

3.3.1.1 Average Local Saliency

The average local saliency is calculated as the average of pixel-wise saliency values over the proposed region. As described previously in Section 3.1, regions with high saliency have a high probability of belonging to an object. Therefore, the sum of saliency values of pixels of a region can be used as an objectness cue. In order to remove the bias for large segments, we normalize the sum to the area of the region to obtain an average local saliency value for the region being scored as given in (30). This scoring ensures that the segments resulting from under-segmentations of the image are penalized. However, it is prone to over-segmentations i.e. region proposals that are wholly contained inside an object tend to have a very high average local saliency value.

$$\Theta_{Saliency} = \frac{\sum_p \Phi(p)}{Area} \tag{30}$$

Figure 28 provides an example of this. The saliency cue is successful in penalizing segments that contain some part of the background in addition to the object. However, as shown in the image at bottom right corner, the measure provides very high score to regions that form a part of the interior of the object.

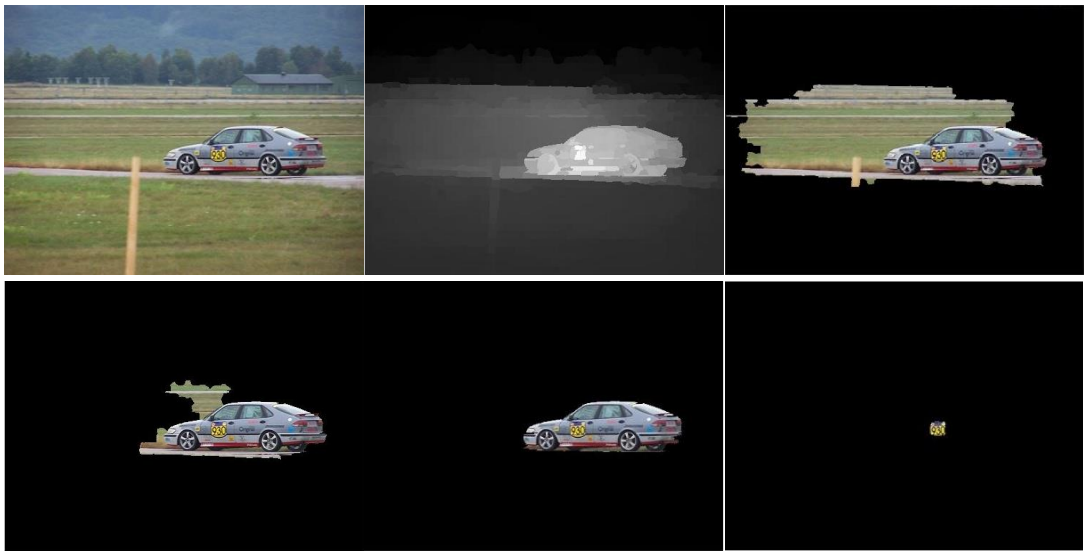

Figure 28. *Average Local Saliency: (from left to right, top to bottom) Original Image, Saliency Map, Proposals with scores 0.22,0.42,0.62 and 0.91*

3.3.1.2 Perimeter Edge Density

The contour edge density of a segment is calculated by summing up the contour strengths on the perimeter of the segment being scored. Object proposals methods of [8] and [12] use a similar cue in their objectness measures. However, we adapt a comparatively faster

approach by simply summing up edge strengths on the perimeter of the segment. This compliments the average local saliency cue described earlier by penalizing over segmentations as objects tend to have a stronger density of pixels on their boundary as compared to their interior. The edge map is generated using the well-known method of [28]. Moreover, the sum is normalized to the perimeter of the region to remove a bias towards larger regions. The measure takes the form of the following equation:

$$\Theta_{ED} = \frac{\sum_p E(p)}{Perimeter} \tag{31}$$

Figure 29 shows a case where the boundary edge density measure is successful in identifying proposals that provide accurate enclosures of objects. The only drawback of this approach is when objects have complex textures on their surface. The textures are labelled by the edge detector as edges and a segment enclosing them can have a high boundary edge density score.

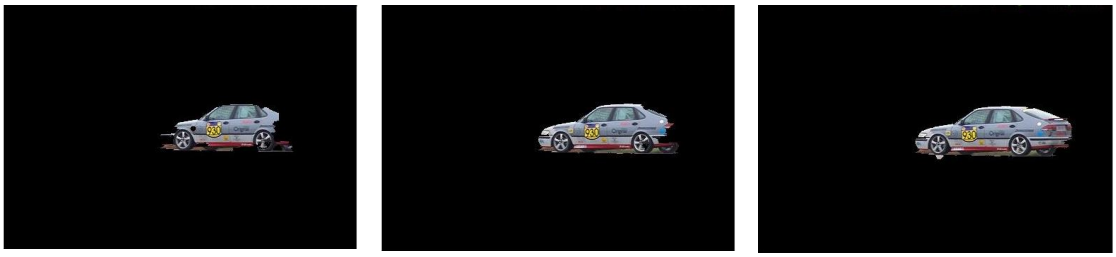

Figure 29. *Proposals with edge density scores 0.23,0.27 and 0.31(from left to right)*

3.3.1.3 Eigenvalues

As described in detail in Section 2.4.2, the optimum labelling vector for the minimization criteria of QCUT is linked with the eigenvector corresponding the minimum eigenvalue of the Hamiltonian matrix. Moreover, as discussed in Section 3.2.1, we further explore other eigenstates as well corresponding to higher eigenvalues. However, as the saliency map corresponding to the ground state provides the most accurate salient object segmentation, proposals extracted from it are most likely to belong to an object of interest. From this, we can coarsely infer a negative correlation between the eigenvalue of the saliency map that is used to generate the proposal and its objectness. We use the reciprocal of the square-root of eigenvalue as an objectness score. In case of superposition, we take the mean eigenvalue of the eigenvectors being superposed. The mathematical form of the eigenvalue based score is as follows:

$$\Theta_{Eigenvalue} = \frac{1}{\sqrt{\lambda_n}} \tag{32}$$

where the saliency map used to generate the proposal corresponds to the eigenvector of the n_{th} smallest eigenvalue. Figure 30 shows saliency maps of an image corresponding

to progressively higher eigenvalues. It can be seen from this that saliency maps corresponding to higher eigenvalues are more prone to producing inaccurate proposals.

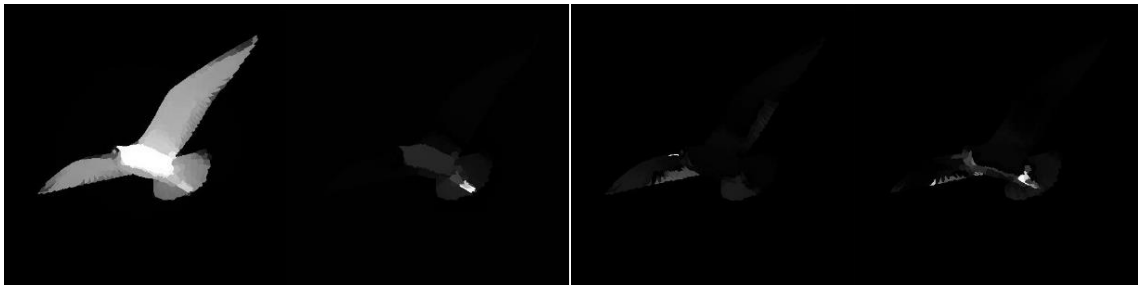

Figure 30. *Saliency maps corresponding to successively larger eigenvalues*

However, it must be kept in mind that as shown in Section 3.2, this does not hold true for all the cases and in some images, eigenvectors corresponding to a higher eigenvalue can produce saliency maps that highlight the object of interest better than the ground state eigenvector.

3.3.1.4 Compactness

Compactness is an objectness measure based on the shape of the segment being scored. It is defined as the ratio of perimeter to the area of the segment. The objective here is to penalize segments having an irregular shape and indefinite form-factor. The mathematical formulation is as follows:

$$\Theta_{Compactness} = \frac{Area}{Perimeter} \tag{33}$$

The rationale behind this measure can be explained using a visual example as shown in Figure 31. Given is an example of three segments along with their compactness scores. From a geometrical viewpoint and knowing nothing about the image that these proposals belong to, we can intuitively infer that the second and third proposals are most probable to belong to a real-life object. The compactness scores are consistent with this as the mask shown in first image has a compactness score of 5.04 whereas those of second and third image have scores 45.44 and 30.1 respectively.

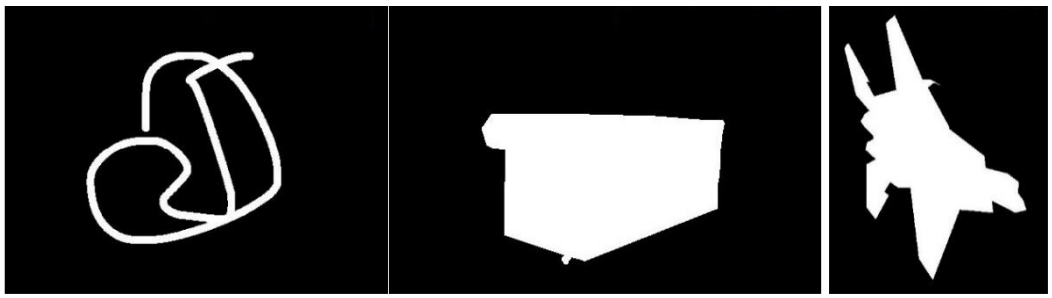

Figure 31. *Compactness scores (left to right): 5.04, 45.44 and 30.1*

Our final scoring function takes the following form:

$$\Theta = \Theta_{ED} * \Theta_{Saliency} * \Theta_{Compactness} * \Theta_{Eigenvalue} \tag{34}$$

Each of the four cues are normalized between 0 and 1 before multiplying them together. The proposals are then ranked in the descending order of the score Θ. Such formulation is mainly adopted for its simplicity and parameter-free nature.

3.3.2 Non-Maximum Suppression (NMS)

Once the proposals are ranked based on the objectness measure described in the previous section, they are subjected to NMS operation, so that the spatial redundancy between them can be minimized. In order to achieve this, we need to define a measure of redundancy or overlap between proposals which is both fast to calculate and provides accurately models the similarity of segments.

To this end, we use the Intersection-over-Union (IoU) overlap measure to quantify the spatial similarity between two proposals. Given two proposals A and B, the IoU overlap is defined as:

$$Overlap(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{35}$$

where $|A \cap B|$ represents the number of pixels common to both the proposals and $|A \cup B|$ is the number of pixels in the union of the two proposals. In statistics, this is also known as the *Jaccard similarity coefficient* and is an effective measure of defining the similarity between two sets [53]. Figure 32 demonstrates the efficacy of IoU overlap for our particular application by showing multiple examples of a pair of proposals and their corresponding IoU values. We can see from the images that the IoU measure is very efficient in correctly scoring the overlap or similarity between proposals.

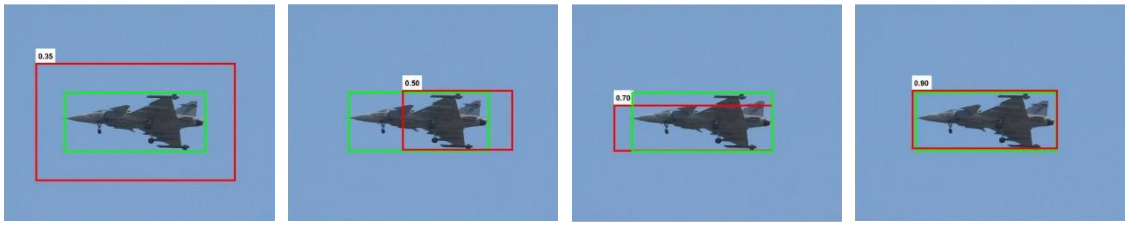

Figure 32. *IoU Overlap: (from left to right) IoU Overlap of 0.35,0.50,0.70 and 0.90 respectively between green and red proposals*

Our non-maximum suppression proceeds as follows. Given a ranked list of proposals $\boldsymbol{r} = \{r_1, r_2, r_3, \dots, r_k\}$, the overlap between two proposals $o(r_i, r_j)$ is given as in (36)

$$o(r_i, r_j) = \frac{|r_i \cup r_j|}{|r_i \cap r_j|} \tag{36}$$

Now the list of ranked proposals is filtered iteratively until the redundancy is minimized according to a predefined criterion. At each iteration i, a proposal r_i is chosen and $\boldsymbol{r}$ is filtered as follows:

$$\boldsymbol{r} \rightarrow \{\boldsymbol{r} - \boldsymbol{s}\} \mid \forall s \in \boldsymbol{s} : o(r_i, s) > \kappa \tag{37}$$

The value of i is incremented at each pass and the process is repeated until $i = |\boldsymbol{r}|$. The threshold κ is a predefined parameter which defines the lower limit for the amount of redundancy that will be tolerated. For instance, setting a value of κ at 0.95 will ensure that no pair of proposals in the filtered list will have an IoU overlap greater than 0.95. Later in Section 0, it is observed that controlling the value of κ allows the optimization of proposals for specific design constraints, which is consistent with the findings of [18]. Moreover, it can be deduced from the above formulation that the value of κ is directly related to the number of proposals in the filtered list and can therefore be used to tune the size of the output set of proposals.

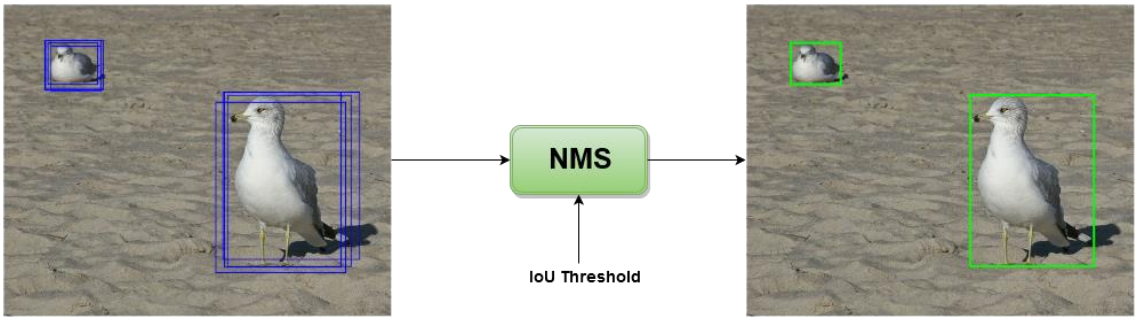

Figure 33. *Non-Maximum Suppression*

Ranking of proposals prior to applying the non-maximum suppression aims to ensure that the loss of recall is minimal. However, given that the number of proposals can reach in the excess of 10^5 for some images, the NMS operation on segments consequently becomes very expensive. To resolve this, we approximate the location of proposals using the tightest fitting bounding boxes and apply the subsequent NMS on the bounding box coordinates. The calculation of overlap in this case simply boils down to calculating areas of two rectangles.

4. EXPERIMENTAL RESULTS

In this section, a comparative analysis is conducted between the proposed method and the contemporary proposal generation methods given in Section 2. We first present a list of performance metrics in Section 4.1 that are generally used to test the methods for their ability to produce accurate proposals and generalize across a variety of object classes [20], [54]. In Section 4.2, we provide the results of various experiments conducted to test the performance of our method against the state of the art.

4.1 Performance Metrics

A brief explanation of each of the performance measure employed when comparing object proposal generation methods is presented next.

4.1.1 IoU

This is the same representation of overlap as introduced in Section 3.3.2 and defined by (43) where it was used to reduce inter-proposal redundancy. For evaluation purpose, the IoU value is calculated between all of the proposals and each of the ground truth annotations. The term IoU threshold implies considering only those proposals which exhibit a predefined minimum IoU overlap with the ground truth objects.

4.1.2 Maximum Achievable Performance

The maximum achievable performance measure is an efficient measure for gauging the peak performance of a method. Given a pool of proposals, the best IoU with each of the ground truth annotated object instances is calculated and averaged.

4.1.3 Recall

Recall is a statistical measure employed to evaluate the results of information retrieval experiments. It is formulated as in (38) and quantifies the fraction of relevant instances that are retrieved.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{38}$$

In the context of object proposal evaluation, recall is the fraction of proposals which have an IoU above a certain fixed value with ground truth.

4.1.4 Average Best Overlap (ABO)

Average best overlap (ABO) is an important measure for measuring the class-specific performance of the proposals. Given the list of proposals $\boldsymbol{r}$ and the ground truth annotations $\boldsymbol{g}$ where $\boldsymbol{g^j} \subset \boldsymbol{g}$ is a set of annotations belonging to a single category $c_j \in \boldsymbol{C}$, the average best overlap for the class c_j is formulated as below:

$$ABO(\boldsymbol{r}, c_j) = \frac{1}{|\boldsymbol{g^j}|} \sum_{g_i^j \in \boldsymbol{g^j}} \max\ (overlap(r_i, g_i^j)\ \forall\ r_i \in \boldsymbol{r}) \tag{39}$$

Another similar measure used is the mean average best overlap (MABO) which is the average of ABO measures over all classes. It is formulated as below:

$$MABO(\boldsymbol{r}) = \frac{1}{|\boldsymbol{C}|} \sum_{c_j \in C} ABO(\boldsymbol{r}, c_j) \tag{40}$$

4.2 Evaluation

4.2.1 Preliminary Evaluation

In order to determine the best proposal extraction strategy from the ones proposed in Section 3.1, we have evaluated the proposals generated using each of the methods presented. First, the superposition coefficients were learnt as described in Section 3.2.2.1 using annotations from PASCAL VOC 2012 [36] training dataset. It consists of 1464 images with 3507 annotated objects which are utilized to provide coefficients for 3507 unique superpositions.

For evaluation, the number of proposals generated by each method was kept almost identical by controlling the number of superpositions used. In case of adaptive thresholding, saliency maps corresponding to 25 smallest eigenvectors and all 3507 unique superpositions of 10 smallest eigenstates were used. For multilevel thresholding, as the number of proposals per saliency map is comparatively higher, saliency maps corresponding to 25 eigenvectors and 25 unique superpositions were used. The coefficients for these superpositions were obtained by clustering the set of coefficients obtained from the training dataset using k-medoids algorithm [55]. Moreover, none of the generated proposals were suppressed in order to get an accurate estimate of the maximum achievable quality of the pool of proposals.

The evaluation was carried out by determining the maximum achievable quality in terms of IoU with the ground truth annotations. The experiments were performed on PASCAL

VOC 2012 [36] validation dataset which has 1449 images and pixel-wise ground truth annotations available ofor3422 objects in those images.

Table 1. Maximum achievable quality of different proposal extraction methods

	Number of Proposals	**Maximum Achievable Quality**
Adaptive Thresholding	10671	0.6564
Multi-level Thresholding	10035	0.7505

Table 1 chronicles the findings of this experiment. As shown, multi-level thresholding achieves a higher maximum achievable quality in fewer number of proposals. In light of this, multi-level thresholding was selected as the proposal extraction technique and employed in all the subsequent experiments.

4.2.2 Comparison with the state-of-the-art

We benchmark our method by comparing its results with those of other methods across a variety of different experiments. The recall experiments are conducted using the toolbox provided by [20], [54]. Furthermore, to ensure a fair comparison, we evaluate all methods as bounding boxes. The proposed segments from region-based methods are transformed by enclosing the tightest fitting bounding box around them. It can be argued that this transformation has little impact on the performance as an accurate region proposal will always correspond to an accurate bounding box. However, bounding boxes cannot be converted into regions so it would not be possible to evaluate the bounding box based methods of [8], [11], [18], [19] in a segment-based evaluation. Moreover, the object detection methods [27],[38], [39] etc. for which object proposals are primarily used are all based on rectangular windows. Therefore, we have chosen to compare the performance of proposals using bounding box based evaluation. It is important to note that this evaluation still favors bounding-box based methods as they can get a similar score as compared to region based methods without providing precise localizations for the objects. However, in such a case, proposals from a region based method will still be preferred because of their higher localization power which makes them suitable for applications such as robotics.

Experiments presented in the rest of this section are performed on the PASCAL VOC 2007 [25] test set which consists of 4952 images and 14,976 instances of objects belonging to 20 categories. For our method, we use saliency maps obtained from eigenvectors corresponding to 50 smallest eigenvalues. and another 500 superpositions of the 10 lowest energy eigenstates. The superposition coefficients are learnt on PASCAL VOC 2012 [36] training dataset as described in Section 4.2.1 and clustered using k-medoids clustering algorithm [55]. Moreover, we test three variants of our method corresponding to three different thresholds of NMS (κ). Table 2 shows the amount of suppression achieved for

each of the three NMS thresholds and the consequent loss in recall. An NMS threshold of 1 implies that only duplicates are suppressed.

Table 2. Effect of NMS threshold on the number of proposals and average recall

NMS Threshold (κ)	Number of Proposals	Average Recall
1	22920	0.6986
0.95	7682	0.6652
0.85	5289	0.6364
0.75	3200	0.5728

4.2.2.1 Maximum Achievable Performance

For this experiment, the maximum achievable performance is calculated across 9 different thresholds of the number of proposals which aids in assessing the quality of the ranking methodology. For methods that do not incorporate an explicit objectness measure like [14], [16], [17], the performance for lower number of proposals is expectedly worse as compared to those methods that rank the proposals based on a learnt objectness score like [8], [18], [30]. The results of this experiment are shown in Figure 34.

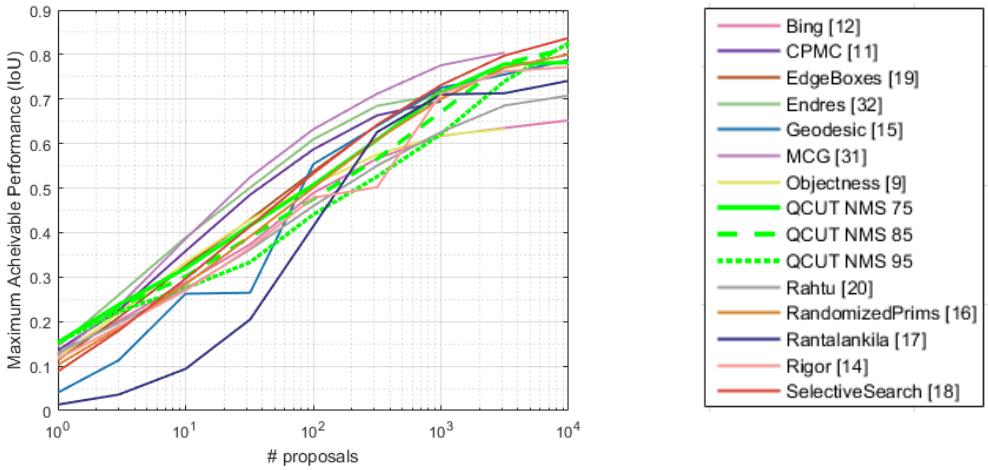

Figure 34. *Maximum Achievable Performance vs number of proposals*

The variant of our method with NMS threshold set at 0.95 achieves a maximum quality of 0.8279 and comes at a close second behind Selective Search's [17] figure of 0.8372. For the top ranked proposal of each method, we achieve the best performance. Moreover, for lower number of proposals, the variants with lower NMS thresholds perform better. This proves that the NMS operation, guided by the scoring, keeps the loss of performance to a minimum while significantly reducing the number of proposals.

4.2.2.2 Recall vs Number of Proposals

This experiment aims to find the number of proposals needed by a method to achieve a particular recall. The threshold for IoU overlap with the ground truth is fixed and only

those proposals are considered as true positives for which the IoU with ground truth is greater than the fixed threshold. The calculation is repeated with different number of top-ranked proposals for each method. The results are shown in Figure 35.

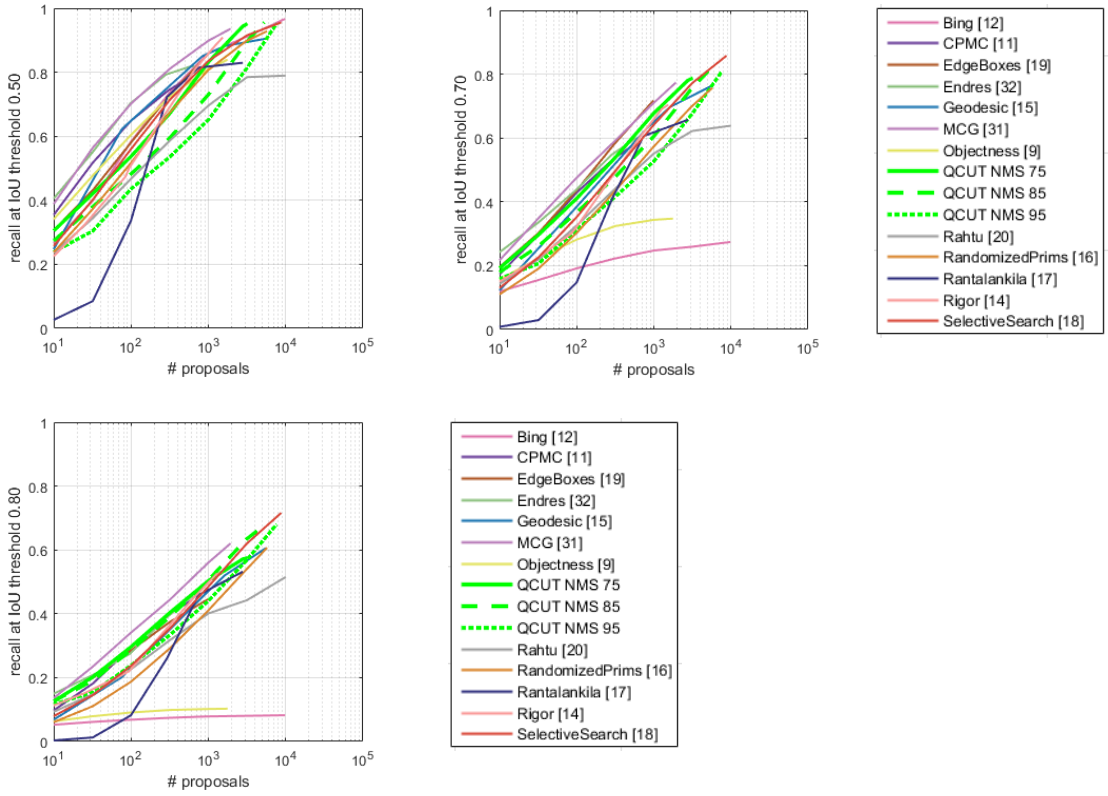

Figure 35. *Recall vs number of proposals at different IoU thresholds*

We can see from the curves that at all IoU thresholds, we achieve competitive results. Moreover, as compared to other methods, the drop in recall at higher IoU values is not as drastic. At a challenging IoU of 0.8, we achieve the second-best results at all number of proposals. This verifies that our method produces highly localized proposals. Also, it can be observed from the curves that for lower IoU thresholds, the variant with lower NMS thresholds perform better. This shows the effectiveness of the scoring-based NMS operation which enables us to optimize the performance at a particular IoU threshold value by achieving similar recall in significantly less number of proposals.

4.2.2.3 Recall vs IoU

The recall vs IoU curves provide complimentary information to the curves presented in the previous section. Here, the number of proposals is fixed and recall is calculated at different IoU thresholds. The curves with number of proposals fixed at 100,1000 and 10,000 are shown in Figure 36. We can clearly see that the curves belonging to bounding box based methods fall quite rapidly at higher overlaps which is a testament to the fact

that region based methods provide better localization as compared to window-based methods.

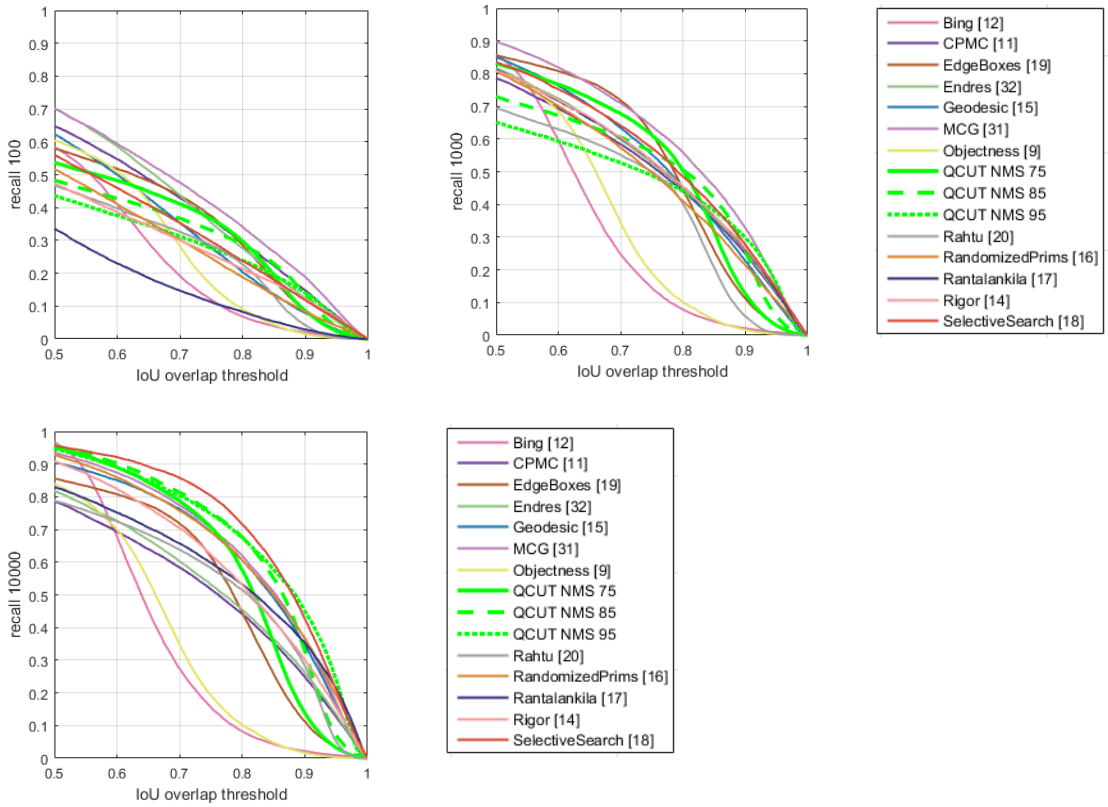

Figure 36. *Recall vs IoU thresholds at fixed number of proposals*

For all the three thresholds for number of proposals, at least one variant of our method achieves competitive results as compared to the state of the art. For the case of 100 and 1000 proposals, the variant with the least NMS threshold (QCUT NMS 75) performs best as it trades-off highly localized proposals to achieve a greater reduction in number of proposals. Given enough number of proposals though, as in the case with curve for 10000 proposals, the variant with the highest threshold performs best as it favors retention of quality proposals over reduction in the total number of proposals.

4.2.2.4 Mean Best IoU vs Object Size

This experiment measures the average best IoU with ground truth objects of different sizes. The motivation behind this is to identify a method's potential susceptibility to objects of abnormally small or large sizes. To conduct this study, the sizes of all annotated ground truth bounding boxes in the PASCAL VOC 2007 [25] test set were calculated. and normalized between 0 and 1. These values are then divided into 10 bins and for each bin, the average maximum IoU with the ground truth is measured for each method. Figure 37 shows the results of this experiment.

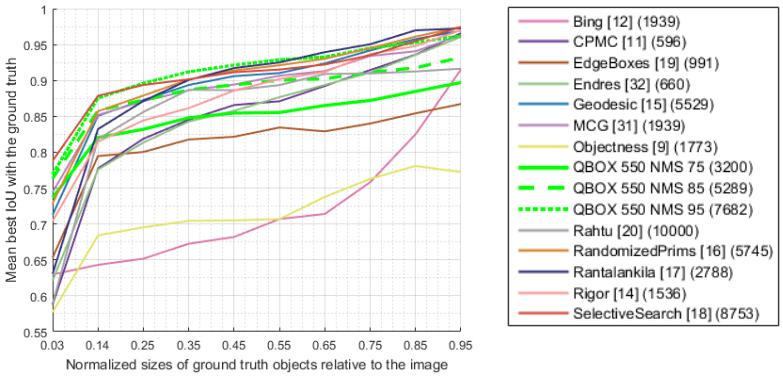

Figure 37. *Mean best IoU with ground truth vs normalized sizes of objects*

We can see from the curves shown in Figure 37 that performance of all methods falls when dealing with objects of very small sizes. However, the performance of all the three variants of our method is shown to be quite impervious to the size of the object and the loss in performance for smaller objects is much lower as compared to other methods. Our worst performing variant still achieves an average IoU of 0.73 with ground truth for the smallest objects in the dataset.

4.2.2.5 Mean Best IoU vs Object Location

In this experiment, the performance is measured against various positions of objects in the image. This stems from the premise that many proposal generation methods suffer a loss of performance when dealing with objects near the image boundaries. To measure this, the normalized distance of an annotated object from the center of the image is measured. The distances for all object instances are then divided into 10 bins and for each bin, the average maximum IoU with the ground truth is calculated. The results of this experiment are shown in Figure 38.

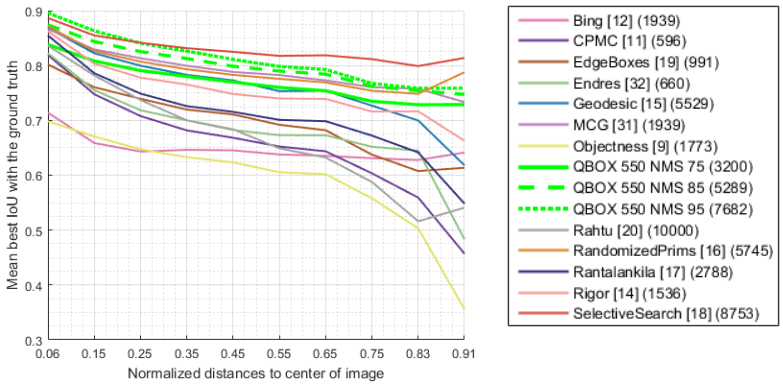

Figure 38. *Mean best IoU with ground truth vs normalized distance to center of images*

As a general trend, we can see that most methods suffer a substantial drop in performance for objects located near the boundary of the image. However, our method is proven to be much more robust to this variation and the loss of performance is very small as compared to other methods.

4.2.2.6 ABO/MABO

The ABO and MABO measures quantify the performance of a method across different classes. This measure is very important in measuring category-independence of the proposals. A high ABO across all classes (which also results in a high MABO), hints towards strong category independence of the proposal generation method. Table 3 shows the ABO values for all the methods against each of the 20 annotated classes in the PASCAL VOC 2007 [25] test set. For each class, the top three methods are highlighted in red, green and blue color respectively. Moreover, for each method, the maximum number of proposals produced are used. The number is shown in parenthesis following the methods' names in Table 3.

Table 3. Average best overlap for 20 annotated classes of PASCAL VOC 2007 test set

	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable
Bing (1939)	0,662	0,652	0,643	0,632	0,631	0,659	0,648	0,706	0,631	0,648	0,685
CPMC (596)	0,743	0,707	0,677	0,577	0,501	0,800	0,685	0,892	0,665	0,734	0,796
EdgeBoxes (991)	0,760	0,780	0,728	0,677	0,576	0,804	0,704	0,822	0,698	0,757	0,776
Endres (660)	0,690	0,786	0,662	0,563	0,536	0,820	0,731	0,880	0,685	0,725	0,835
Geodesic (5529)	0,759	0,830	0,747	0,668	0,638	0,874	0,787	0,920	0,784	0,796	0,881
MCG (1939)	0,813	0,822	0,765	0,718	0,688	0,870	0,814	0,899	0,796	0,820	0,875
Objectness (1773)	0,665	0,664	0,615	0,592	0,539	0,707	0,618	0,727	0,602	0,628	0,720
QBOX 550 NMS 75 (3200)	0,826	0,796	0,797	0,742	0,649	0,828	0,793	0,863	0,762	0,821	0,834
QBOX 550 NMS 85 (5289)	0,858	0,830	0,827	0,768	0,669	0,862	0,822	0,904	0,793	0,851	0,870
QBOX 550 NMS 95 (7682)	0,879	0,843	0,845	0,776	0,657	0,884	0,835	0,933	0,796	0,872	0,893
Rahtu (10000)	0,780	0,762	0,694	0,627	0,517	0,812	0,663	0,879	0,623	0,711	0,828
RandomizedPrims (5745)	0,854	0,834	0,786	0,733	0,662	0,873	0,794	0,919	0,800	0,829	0,887
Rantalankila (2788)	0,763	0,758	0,707	0,582	0,566	0,834	0,716	0,928	0,737	0,750	0,874
Rigor (1536)	0,773	0,799	0,746	0,687	0,637	0,860	0,789	0,917	0,762	0,798	0,824
SelectiveSearch (8753)	0,865	0,869	0,829	0,765	0,727	0,886	0,830	0,928	0,847	0,854	0,904

	dog	horse	motorbike	person	ɔottedplan	sheep	sofa	train	tvmonitor	MABO
Bing (1939)	0,677	0,655	0,656	0,652	0,640	0,645	0,693	0,686	0,646	0,657
CPMC (596)	0,866	0,783	0,751	0,660	0,641	0,699	0,881	0,819	0,786	0,733
EdgeBoxes (991)	0,826	0,798	0,781	0,699	0,691	0,749	0,812	0,796	0,788	0,751
Endres (660)	0,861	0,799	0,809	0,680	0,657	0,680	0,891	0,853	0,753	0,745
Geodesic (5529)	0,897	0,853	0,842	0,767	0,763	0,772	0,915	0,876	0,845	0,811
MCG (1939)	0,884	0,841	0,834	0,788	0,765	0,791	0,917	0,872	0,860	0,822
Objectness (1773)	0,707	0,689	0,663	0,621	0,610	0,598	0,736	0,710	0,640	0,653
QBOX 550 NMS 75 (3200)	0,860	0,826	0,808	0,761	0,778	0,803	0,860	0,833	0,816	0,803
QBOX 550 NMS 85 (5289)	0,901	0,864	0,842	0,792	0,810	0,836	0,902	0,877	0,849	0,836
QBOX 550 NMS 95 (7682)	0,928	0,886	0,860	0,801	0,816	0,849	0,922	0,899	0,866	0,852
Rahtu (10000)	0,854	0,830	0,783	0,695	0,647	0,696	0,866	0,863	0,766	0,745
RandomizedPrims (5745)	0,899	0,840	0,844	0,771	0,757	0,804	0,932	0,875	0,868	0,828
Rantalankila (2788)	0,901	0,835	0,788	0,713	0,687	0,713	0,928	0,865	0,824	0,773
Rigor (1536)	0,887	0,833	0,814	0,739	0,732	0,782	0,904	0,855	0,847	0,799
SelectiveSearch (8753)	0,920	0,862	0,872	0,814	0,815	0,836	0,935	0,894	0,901	0,858

We can see that at least one of our variant is consistently ranked among the top three methods for all classes. Moreover, two out of the three variants are in the top three methods ranked based on MABO scores. We achieve a maximum MABO of 0,852 which is a testament to the category independence of our framework.

5. CONCLUSION

In this work, we have proposed a novel category independent framework for generation of object proposals that hinges on a unique relationship between visual saliency estimation and quantum mechanical principles. A review of the current state-of-the-art in proposal generation reveals that all methods incorporate a degree of supervised learning from ground truth annotations of images. Our research aimed at exploring an unsupervised strategy that is completely devoid of any learnt notions of objectness and is therefore, class-agnostic by design. The competitive results on evaluation benchmarks provide an interesting research insight that computer vision algorithms can be designed to find objects in digital images without having any prior information about the objects.

We found that a saliency estimation technique provided fairly accurate models of visual attention. However, many objects in natural scenes, in-spite of their distinct characteristics, are not prominent and require searching of visual space. Therefore, the method was modified to 'look' at other non-salient parts of the image and has been shown to efficiently perform the task of category independent object localization in images. We found the saliency estimation method of Quantum Cut to be uniquely suited for this purpose. Building upon the proposed parallelism between spectral clustering and quantum mechanics, we found that multiple eigenstates of the quantum system can be used to extend the search for objects in the image beyond the salient ones. We also established that the principle of superposition can be exploited by linearly superposing eigenstates of the quantum system to further diversify the search of objects. This is the first time that the principle of quantum superposition has been employed in image processing applications.

Evaluation of our method against contemporary methods provides useful insight towards its utility. The maximum achievable performance of our proposed method is 0.8259 which is only slightly less than the top performing method's figure of 0.8372. Furthermore, we achieve a MABO score greater than 80% across all classes which is even better than methods that explicitly learn objectness measures across the same classes. Moreover, the proposed NMS strategy based on the unsupervised scoring function alleviates redundancy by reducing the number of proposals by 300% while keeping the loss of recall to less than 0.03. The only drawback of the proposed strategy is the large number of proposals required to achieve the high performance on the evaluated benchmarks. The NMS operation provides a partial resolution to this problem but the number of proposals required is still high as compared to some of the other methods. A possible reason for this can be that as our method encodes no information about the objects in the dataset, it might be producing proposals that localize objects in the image which are not annotated at all.

To conclude, the research work involved using visual saliency estimation for generating class-independent object proposals. By taking inspiration from the way human visual system detects objects in natural scene, the visual saliency estimates were extended to extend the search of objects. Moreover, diversification of object proposals was achieved by exploiting, for the first time, the principle of quantum superposition in the realm of image processing. Evaluation results showed that the proposed method achieves very competitive results without employing any form of supervised learning. This opens up a very research direction which would enable computer algorithms of the future to detect and interact with objects that they have never encountered before.

6. REFERENCES

[1] R. Hess and D. Field, “Integration of contours: New insights,” *Trends Cogn. Sci.*, vol. 3, no. 12, pp. 480–486, 1999.

[2] G. Taylor, D. Hipp, A. Moser, K. Dickerson, and P. Gerhardstein, “The development of contour processing: Evidence from physiology and psychophysics,” *Front. Psychol.*, vol. 5, no. JUL, pp. 1–10, 2014.

[3] A. Treisman and G. Gelade, “A feature integration theory of attention,” *Cogn. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[4] P. Viola and M. Jones, “Robust Real-Time Face Detection,” *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[5] M. Darms, C. R. Baker, P. E. Rybski, and C. Urmson, “Vehicle detection and tracking for the Urban Challenge Vehicle Detection and Tracking for the Urban Challenge,” pp. 57–67, 2008.

[6] M. Field, A. Bruce, and R. Land, “Real-Time Face Detection and Tracking,” no. December, 2012.

[7] D. Forsyth *et al.*, “Finding pictures of objects in large collections of images,” *Object Represent. Comput. Vis. II*, pp. 335–360, 1996.

[8] B. Alexe, T. Deselaers, and V. Ferrari, “What is an object ?”

[9] P. Arbel, J. T. Barron, and U. Polit, “Multiscale Combinatorial Grouping,” vol. 500.

[10] J. Carreira and C. Sminchisescu, “CPMC: Automatic object segmentation using constrained parametric min-cuts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1312–1328, 2012.

[11] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, “BING: Binarized Normed Gradients for Objectness Estimation at 300fps.”

[12] I. Endres and D. Hoiem, “Category Independent Object Proposals.”

[13] A. Humayun, F. Li, and J. M. Rehg, “RIGOR: Reusing inference in graph cuts for generating object regions,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 336–343, 2014.

[14] P. Kr, “Geodesic Object Proposals.”

[15] S. Manen, “Object Proposals with Randomized Prim â€™ s Algorithm,” 2013.

[16] P. Rantalankila, J. Kannala, and E. Rahtu, “Generating object segmentation proposals using global and local search,” *Cvpr*, pp. 2417–2424, 2014.

[17] K. E. A. Van De Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders, “Segmentation as Selective Search for Object Recognition.”

[18] C. L. Zitnick and P. Doll, “Edge Boxes:Locating Object Proposals from Edges,” *Eur. Conf. Comput. Vis.*, pp. 1–15, 2014.

[19] E. Rahtu, J. Kannala, and M. Blaschko, “Learning a Category Independent Object Detection Cascade,” pp. 1052–1059, 2007.

[20] J. Hosang, R. Benenson, P. Dollár, and B. Schiele, “What makes for effective detection proposals?,” Feb. 2015.

[21] X. Hou and L. Zhang, “Saliency Detection: {{A}} Spectral Residual Approach,” *Comput. {{Vision}} {{Pattern Recognition}}, 2007. {{CVPR}}’07. {{IEEE Conf.*, no. 800, pp. 1–8, 2007.

[22] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Trans. {PAMI}*, vol. 8, no. 6, pp. 679–698, 1986.

[23] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, “Using Multiple Segmentations to Discover Objects and their Extent in Image Collections,” *2006 {IEEE} Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, pp. 1605–1614, 2006.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[25] M. Everingham and J. Winn, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Development Kit,” *Challenge*, vol. 2007, pp. 1–23, 2007.

[26] N. Dalai, B. Triggs, I. Rhone-Alps, and F. Montbonnot, “Histograms of oriented gradients for human detection,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 2005. CVPR 2005*, vol. 1, p. 0, 2005.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, 2014.

[28] P. Dollár and C. L. Zitnick, “Fast Edge Detection Using Structured Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, 2014.

[29] P. Dollar and C. L. Zitnick, “Structured Forests for Fast Edge Detection,” *2013 IEEE Int. Conf. Comput. Vis.*, pp. 1841–1848, 2013.

[30] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, “Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation,” pp. 1–14, 2015.

[31] I. Endres and D. Hoiem, “Category Independent Object Proposals.”

[32] M. Maire, P. Arbeláez, C. Fowlkes, and J. Malik, “Using contours to detect and

localize junctions in natural images," *26th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR*, 2008.

[33] D. S. Hochbaum, "The Pseudoflow Algorithm: A New Algorithm for the Maximum-Flow Problem," *Oper. Res.*, vol. 56, no. 4, pp. 992–1009, 2008.

[34] S. Wang and J. M. Siskind, "Image segmentation with ratio cut," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 6, pp. 675–690, 2003.

[35] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335–336.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," *httpwwwpascalnetworkorgchallengesVOCvoc2009workshopindexhtml*, vol. 2012, pp. 1–45, 2012.

[37] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[38] R. Girshick, "Fast R-CNN," 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[40] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 34, no. 11, pp. 2274–2282, 2011.

[41] R. C. Prim, "Shortest connection networks and some generalizations.," *Bell Syst. Technol. J.*, 1957.

[42] D. Hoiem, A. N. Stein, and A. a Efros, "Recovering Occlusion Boundaries from a Single Image Recovering Occlusion Boundaries from a Single Image," *Comput. Vision, 2007. ICCV 2007. IEEE 11th Int. Conf.*, no. Figure 2, pp. 1–8, 2007.

[43] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, p. 309, 2004.

[44] M. Stoer and F. Wagner, "A simple min-cut algorithm," *J. ACM*, vol. 44, no. 4, pp. 585–591, Jul. 1997.

[45] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation Normalized Cuts and Image Segmentation," vol. 22, no. March, pp. 888–905, 2005.

[46] Ç. Aytekin, E. C. Ozan, S. Kiranyaz, and M. Gabbouj, "Extended quantum cuts for unsupervised salient object extraction," *Multimed. Tools Appl.*, 2016.

[47] M. Planck and U. Von Luxburg, “A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering,” *Stat. Comput.*, vol. 17, no. March, pp. 395–416, 2006.

[48] C. Aytekin, E. C. Ozan, S. Kiranyaz, and M. Gabbouj, “Visual saliency by extended quantum cuts,” *Proc. - Int. Conf. Image Process. ICIP*, vol. 2015–Decem, no. November 2016, pp. 1692–1696, 2015.

[49] C. Aytekin, S. Kiranyaz, and M. Gabbouj, “Automatic object segmentation by quantum cuts,” *Proc. - Int. Conf. Pattern Recognit.*, pp. 112–117, 2014.

[50] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient Object Detection: A Benchmark,” *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[51] Ç. Aytekin, S. Kiranyaz, and M. Gabbouj, “Learning to rank salient segments extracted by multispectral Quantum Cuts,” *Pattern Recognit. Lett.*, vol. 72, pp. 91–99, 2015.

[52] N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” *IEEE Trans. Syst. Man. Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[53] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, “Using of Jaccard Coefficient for Keywords Similarity,” *Int. MultiConference Eng. Comput. Sci.*, vol. I, pp. 380–384, 2013.

[54] J. Hosang, R. Benenson, and B. Schiele, “How good are detection proposals, really?,” *Bmvc*, pp. 1–25, 2014.

[55] L. Kaufman and P. J. Rousseeuw, “Clustering by means of Medoids,” in *Statistical Data Analysis Based on the L1–Norm and Related Methods*, Y. Dodge and North-Holland, Eds. 1987, pp. 405–416.