UNIVERSITY of CALIFORNIA
Santa Barbara

# Modeling Eye Tracking Data with Application to Object Detection

A dissertation submitted in partial satisfaction of the
requirements for the degree

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Karthikeyan Shanmuga Vadivel

Committee in charge:

Professor B. S. Manjunath, Chair
Professor Kenneth Rose
Professor Miguel Eckstein
Professor Shivkumar Chandrasekaran
Professor Scott Grafton

December 2014




The dissertation of Karthikeyan Shanmuga Vadivel is approved.

---

Professor Kenneth Rose

---

Professor Miguel Eckstein

---

Professor Shivkumar Chandrasekaran

---

Professor Scott Grafton

---

Professor B. S. Manjunath, Committee Chair

September 2014

Modeling Eye Tracking Data with Application to Object Detection



by

Karthikeyan Shanmuga Vadivel

To Amma, Appa, Keerthana and Renuka

for all the sacrifices you have made and for inspiring me.

# Acknowledgements

The last six years at UC Santa Barbara have been among the most exciting phases of my life. I got the chance to make several life long friends and travel to new places, and met my wife.

First, I would like to thank Prof. Manjunath for giving me the opportunity to work in his lab during the last five years. I have been lucky to work on several interesting projects in a variety of fields related to computer vision and pattern recognition and it was his vision which enabled me to work on thought provoking problems. He has always been a pillar of support and his patience and positive outlook were vital in getting through challenging times during the journey. Though he has a busy schedule, he has always been very approachable. He has had tremendous impact in shaping my research outlook and the way I approach new problems.

Next, I want to thank my lab-mates, several of them have contributed directly or indirectly to this thesis. Eye tracking features prominently in this thesis and there was considerable groundwork laid by other lab-mates which I benefited from. The eye tracker was purchased as a part of the camera network project and I want to acknowledge Office of Naval Research (ONR) grant which made it possible and the effort put in by the group to set up this infrastructure. The eye tracker was initially used by students who did an internship in summer in Vision Research Lab and their mentors especially Carlos Torres who helped me set up the eye tracker. Further, our research proposal

primarily contributed by Dr. Carter De Leo, Prof. Manjunath and Prof. Grafton was funded by Institute for Collaborative Biotechnologies (ICB) which laid the foundation for me to work on eye tracking enhanced computer vision, a relatively unexplored research area.

In the research projects related to my thesis, several lab-mates actively collaborated in the formulation and experimental analysis of the ideas. I want to thank Dr. Vignesh Jagadeesh for the invaluable discussions which significantly contributed to this thesis. I especially want to thank Thuyen Ngo for his collaboration in my final project. He is very dedicated and efficient and was of great help in the experimental design and obtaining publishable results. I am looking forward to working more with him in the future during my stay at UC Santa Barbara as a post-doctoral researcher. I also want to thank Renuka who has been very involved in brainstorming new ideas, performing experiments and reviewing my articles.

I have been fortunate enough to work on a plethora of problems beyond my thesis during my PhD at UC Santa Barbara. During my first two years, I worked on different problems related to subspace analysis. I closely collaborated with Dr. Swapna Joshi and Prof. Grafton during this time. Dr. Swapna Joshi is a great mentor and friend and she greatly helped me transition from a student to a researcher. We have discussed countless ideas during her PhD at UCSB and she greatly helped me develop technical writing skills and research maturity. I have also collaborated with lab-mates on several

projects. I would like to thank Dr. Diana Delibaltov, Dr. Emre Sargin, Dr. Aruna Jammalamadaka, Dr. Santhoshkumar Sunderrajan, Dr. Thomas Kuo, Utkarsh Gaur and Lakshmanan Nataraj who I directly collaborated in projects with and obtained co-authored publications. I also want to thank Dr. Anindya Sarkar, Dr. Pratim Ghosh, Dr. Luca Bertelli, Dr. Jiejun Xu and Dr. Zefeng Ni for their critical inputs to my projects during my PhD. Additonally I have had several interesting technical discussions with Archith John Bency, Amir Mohaymen, Dmitry Fedorov, Chris Wheat, Niloufar Pourian which gave exposure to new ideas and techniques.

Additionally I want to thank my summer internship mentors - Dr. Felix Fernandes, Dr. Ankur Saxena, Dr. Zhan Ma and Dr. PoLin Lai at Samsung Telecommunications America and Dr. Jiyun Byun and Dr. Ken Sullivan at Mayachitra Inc. - for the opportunity to work on exciting projects. I also want to thank my summer interns Michael Stephens and Taylor Sanchez, who helped implement critical parts of my research projects. I want to thank my dissertation committee members Prof. Scott Grafton, Prof. Kenneth Rose, Prof. Miguel Eckstein and Prof. Shiv Chandrasekaran for their support and brainstorming sessions which significantly contributed to this thesis.

Finally, I want to acknowledge the research grants that supported this work: W911NF-09-0001 from the U.S. Army Research Office, N00014-12-1-0503, NSF III-0808772 and OIA-0941717.

# Curriculum Vitæ

Karthikeyan Shanmuga Vadivel

| | |
|---|---|
| September 2014 | Doctor of Philosophy<br>Department of Electrical and Computer Engineering<br>University of California, Santa Barbara |
| December 2009 | Master of Science<br>Department of Electrical and Computer Engineering<br>University of California, Santa Barbara |
| July 2008 | Bachelor of Technology<br>Department of Electrical Engineering<br>Indian Institute of Technology Madras, India |

**Field of Study**

Computer Vision, Pattern Recognition and Image Processing

**Experience**

| | |
|---|---|
| 2008-2014 | Research/Teaching Assistant<br>University of California, Santa Barbara |
| 2013 | Research Intern<br>Mayachitra Inc., Santa Barbara, CA |
| 2011 | Research Intern<br>Samsung Research Lab, Richardson, TX |

**Publications**

S. Karthikeyan, Vignesh Jagadeesh, Renuka Shenoy, Miguel Eckstein, B.S. Manjunath : From Where and How to What We See, ICCV 2013

S. Karthikeyan, Vignesh Jagadeesh, B.S. Manjunath : Learning bottom up text attention maps for text detection using stroke width transform, ICIP 2013

S. Karthikeyan, Vignesh Jagadeesh, B.S. Manjunath : Learning top down scene context for visual attention modeling in natural images, ICIP 2013

Santhoshkumar Sunderrajan S. Karthikeyan and B.S. Manjunath: Robust Multiple Object Tracking by Detection with interacting Markov Chain Monte Carlo, ICIP 2013

Diana Delibaltov, S. Karthikeyan., Vignesh Jagadeesh, B.S. Manjunath, Robust Biological Image Sequences Analysis using Graph based Approaches, Asilomar Conference on Biological Image Analysis and Microscopy 2012.

S. Karthikeyan, Diana Delibaltov, Utkarsh Gaur, Mei Jiang, David Williams B.S. Manjunath: Unified Probabilistic framework for simultaneous detection and tracking with application to bioimage sequences, ICIP 2012

S. Karthikeyan, Swapna Joshi B.S. Manjunath Scott Grafton: Intra-class Multi output Regression based subspace analysis ICIP 2012

S. Karthikeyan, Felix Fernandes, Zhan Ma, PoLin Lai, Ankur Saxena: Perceptual similarity based robust low-complexity video fingerprinting ICIP 2012

S. Karthikeyan, Mehmet Emre Sargin, Swapna Joshi, B.S. Manjunath, Scott Grafton: Generalized Subspace based High dimensional Density estimation, ICIP 2011

S. Karthikeyan, Utkarsh Gaur, B.S. Manjunath and Scott Grafton: Probabilistic subspace based learning of shape dynamics modes for multi-view action recognition, ICCV Workshop 2013

Lakshmanan Nataraj, S. Karthikeyan, Gregoire Jacob, B.S. Manjunath: Malware Images : Visualization and Automatic Classification, VizSec 2011

E. Alegre, M.T. Garcia-Ordas, V. Gonzalez-Castro, S. Karthikeyan- Vitality assessment of boar sperm using N Concentric Squares Resized (NCSR) texture descriptor in digital images, IBPRIA 2011

Swapna Joshi, S. Karthikeyan, B.S. Manjunath, Scott Grafton, and Kent Kiehl: Anatomical Parts-Based Regression Using Non-Negative Matrix Factorization, CVPR 2010

Vignesh Jagadeesh, S. Karthikeyan and B.S. Manjunath: Spatio-Temporal Optical Flow Statistics (STOFS) for Activity Classi fication, ICVGIP 2010

Aruna Jammalamadaka, Swapna Joshi, S. Karthikeyan, B.S. Manjunath: Discriminative Basis Selection using Non-negative Matrix Factorization, ICPR 2010

Jinal Shah, Rama Kowsalya, Ish Kumar Dham, S. Karthikeyan, Maitreyi Addala, Subham Banerjee, Rahul Maitra, PRP based

flows for AC characterization for TDL coverage validation, Texas Instruments India Technical Conference, 2007

**Patents**

Felix Fernandes, S. Karthikeyan, Zhan Ma, PoLin Lai, Ankur Saxena: Apparatus and Method for Robust Low Complexity Video Fingerprinting, Publication number WO 2013036086 A2

## Abstract

Modeling Eye Tracking Data with Application to Object Detection

by

Karthikeyan Shanmuga Vadivel

This research focuses on enhancing computer vision algorithms using eye tracking and visual saliency. Recent advances in eye tracking device technology have enabled large scale collection of eye tracking data, without affecting viewer experience. As eye tracking data is biased towards high level image and video semantics, it provides a valuable prior for object detection in images and object extraction in videos. We specifically explore the following problems in the thesis: 1) eye tracking and saliency enhanced object detection, 2) eye tracking assisted object extraction in videos, and 3) role of object co-occurrence and camera focus in visual attention modeling.

Since human attention is biased towards faces and text, in the first work we propose an approach to isolate face and text regions in images by analyzing eye tracking data from multiple subjects. Eye tracking data is clustered and region labels are predicted using a Markov random field model. In the second work, we study object extraction in videos using eye tracking prior. We propose an algorithm to extract dominant visual tracks in eye tracking data from multiple subjects by solving a linear assignment

problem. Visual tracks localize object search and we propose a novel mixed graph association framework, inferred by binary integer linear programming. In the final work, we address the problem of predicting where people look in images. We specifically explore the importance of scene context in the form of object co-occurrence and camera focus. The proposed model extracts low-, mid- and high-level and scene context features and uses a regression framework to predict visual attention map. In all the above cases, extensive experimental results show that the proposed methods outperform current state-of-the-art.

# Contents

**List of Tables** xv

**List of Figures** xvi

**1 Introduction** 1
- 1.1 Motivation . . . 2
- 1.2 Overview of proposed methods . . . 6

**2 Eye tracking review** 11
- 2.1 Eye tracking data basics . . . 15
- 2.2 Eye Trackers . . . 18
- 2.3 Eye tracking setup . . . 22
- 2.4 Eye tracking in proposed research . . . 24

**3 Eye tracking enhanced object detection in images** 26
- 3.1 Introduction . . . 28
- 3.2 Faces and Text Eye Tracking Database . . . 31
- 3.3 Faces and Text Localization from Eye Tracking Data . . . 34
- 3.4 Performance of Face and Text Localization . . . 42
- 3.5 Applications . . . 47
- 3.6 Discussion, Summary and Future Work . . . 53

**4 Eye tracking assisted object extraction from videos** 56
- 4.1 Introduction . . . 58
- 4.2 Eye tracking dataset on videos . . . 61
- 4.3 Proposed approach to extract objects from videos using eye tracking prior 62
- 4.4 Experimental results . . . 72
- 4.5 Summary . . . 78

4.6 Future Work . . . . . 79

**5 Saliency enhanced computer vision** **80**
5.1 Introduction . . . . . 81
5.2 Background: Stroke Width Transform . . . . . 84
5.3 The Proposed Approach to learn text attention maps . . . . . 86
5.4 Experiments and Results . . . . . 90
5.5 Summary . . . . . 99

**6 Role of scene and camera context in visual attention modeling** **101**
6.1 Introduction and related work . . . . . 103
6.2 Eye Tracking Dataset on Light Field images . . . . . 107
6.3 Analysis of eye tracking dataset . . . . . 109
6.4 Visual attention model for light field images . . . . . 121
6.5 Experiments and Results . . . . . 129
6.6 Summary and future directions . . . . . 137

**7 Conclusions and Future Work** **140**
7.1 Future Work . . . . . 143

**Bibliography** **146**

# List of Tables


3.1 Indicates performance of cluster and image level face and text detection from the eye tracking samples.We notice that the recall (marked in bold) is high suggesting that the proposed approach seldom misses face and text detections in images. This is achieved at a sufficiently good precision ensuring that this method can be valuable to localize ROI to reduce the search space for computationally expensive face and text detectors. . . . . 46

3.2 Comparison of the performance of the proposed text detector with eye tracking prior and baseline SWT. There is significant gain in the precision ($\sim$37% compared to baseline) for a small loss in recall ($\sim$2%). This results in improved overall F-Measure. . . . . 52

4.1 Comparison of the performance of our multiple object extraction algorithm with active segmentation using fixations [68]. We also selectively compare the performance of different sub-blocks of our model. We notice that both the object extraction module and eye tracking data contribute equally to extract objects which attract visual attention. . . . 73

5.1 Comparison of the performance of our algorithm and SWT . . . . . 96

6.1 Comparison of the performance of Judd et al. [10] with the proposed approach to predict the best 2D image from a Lytro image . . . . . 136

# List of Figures


1.1 Block diagram illustrates the primary theme of this thesis. We want to explore problems where eye tracking data can improve computer vision problems related to object extraction in images and videos. Additionally, we also want to investigate the inverse scenario of how object detection can benefit the prediction of eye tracking data in images. . . . 3

1.2 Figure overlays eye tracking data (shown as green circular dot) from a subject onto video frames from two different video sequences. We notice that eye tracking data in this free viewing task is biased towards objects in the scene and attention shifts from one object to another. . . . 5

1.3 This image shows the eye tracker setup used in the proposed work. Detailed description of the eye tracker setup is discussed in Chapter 2 . 6

1.4 Lists the computer vision algorithms which we aim to improve using the hybrid eye tracking object detection algorithms. (Left) Text detection in natural scenes (Center) Dog and Cat detection (Right) Object extraction in videos . . . . . 7

2.1 Study by Yarbus which indicates the importance of the task which the person is asked to perform on eye movements [102] . . . . . 14

2.2 Example eye tracking data on images during a free viewing task. Eye movement data on images typically consists of alternate fixations and saccades. The fixations are represented by circles and the saccades are represented by lines. The fixations indicate information gathering stage from an image region. The saccades indicate attention shifts from one fixation to another. The entire eye movement sequence is called a visual scanpath. . . . . 17

2.3 Example of electrooculography based eye tracker [16]. . . . . 21

2.4 Illustrates the eye tracker setup using Eyelink 1000 eye tracking device . 23

3.1 Left to right: 1. Input image. 2. Eye Tracking fixation samples from multiple subjects overlaid on the image 3. The eye tracking regions identified by the proposed algorithm as faces (blue) and text (green) 4. The final detection outputs of face and text detector focusing on the priors provided by eye tracking. . . . . 28
3.2 Examples of images from our dataset consisting of text, human faces, dogs, cats and other background objects . . . . . 32
3.3 Shows example of faces and text in two scenarios each. The fixations are marked as red points and saccades as blue lines. Multiple faces in the image where we consistently observe inter-face saccades (top left).A large single face where several saccades are observed in the eyes, nose vicinity (top right). Text with four words where a dense saccadic presence is observed between words (bottom left).A clip from one of the images showing a single word, whose cluster takes a nearly elliptical shape (bottom right). . . . . 34
3.4 Visualizing the text MRF potentials. 1. (Top left) Input image. 2. (Top right) Eye tracking samples overlay 3. (Bottom Left) Clustered eye tracking fixation locations from multiple subjects overlaid on the image 3. (Bottom Right) Visualizing the unary and interaction potentials of the clusters for the text MRF. The unary is color coded as green, the bright values indicating high unary potentials of a cluster belonging to text class. The interaction is marked by the blue lines between clusters, whose thickness is indicative of text-text interaction magnitude. . . . . 40
3.5 Indicates the process to calculate the unary and the pairwise potentials. The unary potentials are computed as $-log$ of posterior probabilities obtained from a quadratic discriminant analysis classifier from intra-cluster features. In similar vein, the pairwise potentials are obtained from inter-cluster features . . . . . 43
3.6 Left: Input image with the ground truth for face (blue) and text (green). Center: Clustered eye tracking data overlay on input image. Right: Face (blue) and text (green) cluster labels propagated from ground truth. 44
3.7 Examples of good face detections from the proposed algorithm. Red fixation points correspond to face and blue corresponds to background. (a) In the presence of salient distracting object (shoe) the face (cat) is reliably detected. (b) We notice that even in challenging scenarios where multiple faces are present, the proposed approach detects reliably. . . . 45

3.8 Examples of good text detections from the proposed algorithm. Red fixation points correspond to text and blue corresponds to background. (a) Text line is reliably detected even in the presence of several other fixations near the region of interest. (b) Text is detected correctly in the presence of more salient object (person face). . . . . 45
3.9 Example scenario where the proposed approach fails to detect face (left) and a text word (right). The eye tracking samples detected as face in (a) and text in (b) are shown in red and the samples detected as background (both (a) and (b)) are indicated in blue. . . . . 46
3.10 Example cat and dog face (blue box) and body (green box) detections from the proposed algorithm. . . . . 48
3.11 Plotting Average Precision (AP) of Cat head (top left) Dog head (top right), Cat Body (bottom left) and Dog Body (bottom right). The proposed (green) and baseline (red) curves are plotted against the detector threshold of deformable parts model. The maximum AP of baseline and proposed algorithm is comparable in all cases, however, the AP of the proposed approach is higher than baseline in high recall scenarios (low detector threshold) for both the head and body detectors of cats and dogs. Therefore, on an average the proposed approach is more stable over the detector threshold parameter than the baseline. . . . . 49
3.12 An example scenario where the head detector of the proposed approach (b) operating only in the attention region (c) marked in blue outperforms the baseline cat head detector (a). The baseline detector has a false detection as noticed in (a). Finally, red points in (c) denotes the cluster identified as face/head from which the blue attention region is constructed. . . . . 50
3.13 Examples of images where the proposed text detection approach performs reliably. . . . . 52
3.14 Two example scenarios ((a)-(c) and (d)-(f)) where SWT results ((a) and (d)) are outperformed by the proposed approach ((b) and (e)). The attention regions ((c) and (f)) shows the eye tracking samples classified as text in red and the ROI used by the text detector in blue. Therefore, as the false positive portion in SWT (red boxes in (a) and (d)) is removed by the generated text attention region, we obtain better detector precision in these images. . . . . 53

4.1 A simple illustration of the proposed problem. Given a video sequence, we collect eye tracking data in the sequence from multiple subjects and utilize this information to extract visually salient objects. . . . . 61

4.2 Illustrates some example frames from the videos in the eye tracking dataset collected from Chen Xiph.org, GaTech and SegTrack datasets. We note that the dataset consists of single and multiple stationary and moving objects with moving and stationary backgrounds. . . . . 63
4.3 Block diagram of the proposed approach to extract multiple objects from videos using eye tracking prior. The top row indicates the eye tracking processing stage. The bottom row is the multiple object extraction framework guided by the visual tracks. . . . . 64
4.4 Illustrates that the saccades in the direction of optical flow are probable object saccades (smooth pursuit) and should be utilized along with the fixations for object localization. Saccades not in the direction of optical flow indicate attention shift from one object to another and can be pruned. 65
4.5 The spatio-temporal graph to extract multiple objects is highlighted here. The temporal costs shown in blue indicate inter-frame cost to connect a path through two bounding boxes in successive frames. The intra-frame spatial costs are indicated in green. They aim to penalize extraction of the same object in multiple paths. . . . . 70
4.6 Shows example results using the proposed approach to extract multiple objects represented by bounding boxes. We see the proposed approach is able to localize different visually salient objects in the video sequences with reasonable accuracy. . . . . 75
4.7 Shows example results from the proposed approach after applying grabcut based video segmentation to the extracted multiple object bounding boxes. We see the proposed approach is able to segment multiple objects in the video sequences with reasonable accuracy. . . . . 76
4.8 Shows some segmentation results using [68]. We notice that in the top row, when the fixation slightly positioned outside the object of interest, [68] breaks down. In addition, the algorithm suffers from similar issues in the bottom row as well as not being robust to the presence of the occluding pole. . . . . 77

5.1 Left to right: 1. Input image. 2. Text attention map derived using visual attention features. 3. The text detection output indicated by the blue rectangle. Best viewed in color. . . . . 82
5.2 Left to right: 1. The input image. 2. Stroke Width Transform Image. 3. Connected components and geometric filtering. 4. Final Detections (blue boxes). Best viewed in color. . . . . 86
5.3 Example of visual attention features computed in an image. . . . . 87
5.4 Block diagram of the training(top) and test (bottom)modules of the visual attention based learning paradigm. . . . . 89

5.5 Left column shows the input image, center column corresponds to human fixation map and the right column illustrates the proposed text attention map. The text attention maps are similar to human fixation map on text centric images. Note that eye fixations only includes foveal or central vision and peripheral vision is not captured. Therefore, as row 1 and 3 only have a single word, eye tracking results are biased towards the center of the word and therefore does not entirely overlap with our text attention map. Moreover, the text attention maps reliably localize the text regions. . . . . 93
5.6 Example detections (blue boxes) in images from ICDAR dataset. Best viewed in color. . . . . 97
5.7 Illustrates two example scenarios where our algorithm (left) outperforms SWT (center). The text attention maps (right) clearly ignores regions where SWT detects false positives. The detections are shown in blue rectangles. Best viewed in color. . . . . 98
5.8 An example image (left) where the proposed algorithm fails and the corresponding attention map (right). In this image the background is very similar to text region, hence, the text attention map fails to localize the text region. The missed detections are shown in red rectangles and the false positive detections in blue rectangles. Best viewed in color. . . 98

6.1 Left-Right. Top-Bottom (1) Image with background person in focus (2) Top 10% attention regions in the image from eye tracking data (3) Image with foreground text in focus (4) Top 10% attention regions in (3). We notice a significant shift in attention across two images of the same scene . . . . . 104
6.2 Plot of focus change with attention change. We notice that there is noticeable correlation (shown by red line) between visual attention change and change in camera focus. The correlation is represented by the best line fir to this data which has a slope of 0.27. . . . . 108
6.3 Examples of images from our dataset consisting of faces, text, people, animals and car etc. at different focus depths. . . . . 111
6.4 Two 2D images from a Lytro image and the corresponding binary focus maps to their right. . . . . 112
6.5 Examples manual object annotations . . . . . 114
6.6 Attention density in 23 manually annotated objects in-focus (blue) and out-of-focus (green). Objects in focus consistently have higher attention density than out-of-focus objects . . . . . 114

6.7 Average initial fixation time in 23 manually annotated objects in-focus (blue) and out-of-focus (red). Objects in focus consistently have lower average time to first fixation compared to out-of-focus objects . . . . . 116
6.8 Visual attention consistency across multiple focus images (Focus ROC curve) . It is shown in reference to the human ROC curve which measures the consistency among fixations in the same image across multiple subjects. . . . . 118
6.9 Row 1 and 2 shows examples of cases where camera focus significantly alters visual attention. We notice categories such as text can have a significant change in the manner in which we view images if their focus attributes change. Rows 3 and 4 highlights examples where camera focus does not significantly alter visual attention. In row 3, the out of focus faces also attract significant attention. In row 4, there is only one salient object in the center of the image and therefore attracts considerable attention irrespective of camera focus. . . . . 119
6.10 The plot shows the consistency of visual attention in the images shown in Fig 6.9 . . . . . 120
6.11 An example high level object layout with N=3 and the corresponding 9 scene context feature maps . . . . . 124
6.12 (a) Has only one salient object (small car) and (b) has one salient object (large). However (c) has both the salient objects. We observe the large face significantly draws attention (green overlay) away from the small car in (c). Best viewed in color. . . . . 125
6.13 Comparison of the performance of the proposed approach with other state-of-the-art visual attention and saliency algorithms using ROC curves. Our approach outerforms other algorithms by a significant margin. However, we still notice considerable gap between machine and human performance. . . . . 131
6.14 Each row contains ground truth attention map (Blue overlay) with a focus setting followed by corresponding predicted visual attention map (Red overlay). The results are shown on two Lytro images (rows) each under two focus settings . . . . . 132
6.15 Weights of scene context features compared to average non-context feature . . . . . 134
6.16 Regression weights of different objects in in-focus and out-of-focus object categories. We notice this plot is similar to the average attention density plot in Figure 6.6 . . . . . 135
6.17 Performance of manual annotations compared to automated detectors. . 135

6.18 Left column indicates the input image. The center column indicates the visual attention map and the right column indicates the region in focus. An analysis of row 1 and 2 indicates that the image in row 1 has higher visual attention in the focused region than image in row 2. In addition we notice that image in row 3 has higher attention in the focused region than image in row 4. Therefore, input image in row 1 and row 3 are preferable to row 2 and row 4 respectively as they capture larger visual attention in the region of focus. . . . . 138

# Chapter 1

# Introduction

"We live in a wonderful world that is full of beauty, charm and adventure. There is no end to the adventures we can have if only we seek them with our eyes open".

*Jawaharlal Nehru*

Recently, there has been significant advancement in eye tracking technology. Current eye trackers have become affordable, accurate and easy to use. These eye trackers do not require the head to be constrained in a specific position and can collect the data in standard viewing postures. This has enabled large scale low-cost availability of such data from multiple subjects especially in multimedia settings. This information can be

extremely beneficial to computer vision algorithms and eye tracking guided computer vision algorithms have become a relevant topic of interest. These hybrid techniques have two design challenges. The first challenge is to effectively identify useful information from the eye tracking data for the task (computer vision problem) one is trying to solve. Secondly, a guiding mechanism needs to be designed that utilizes this eye tracking information to further aid the computer vision task at hand. In addition, in this thesis, we are also interested in predicting eye tracking data given an image. Specifically, we explore the importance of two factors, camera focus and object co-occurrence to improve the prediction of where people look in images.

## 1.1 Motivation

The primary goal of this thesis is to understand the how visual saliency and eye tracking based input can enhance computer vision algorithms and vice versa. Humans are adept at visual tasks and tapping into the contextual information utilized by humans to perform high level tasks can be a valuable tool to develop better computer vision algorithms. In addition, better high level image semantic understanding can benefit algorithms modeling human visual attention, which can help in effective prioritization of information content in images and videos. Figure 1.1 visually illustrates the primary theme of the thesis.

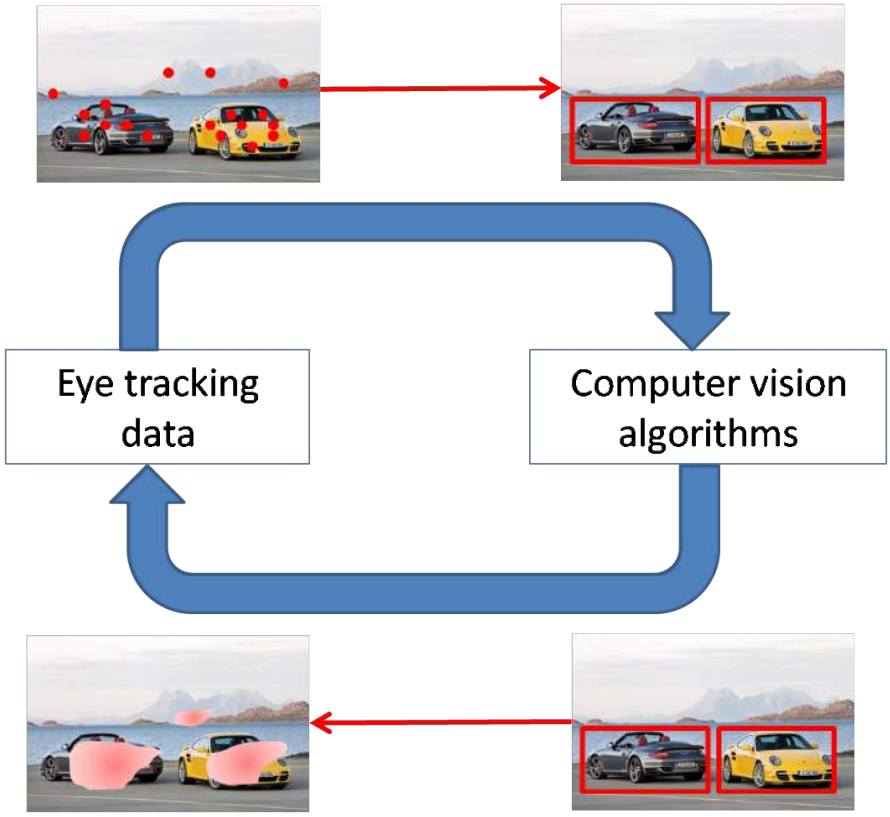



Figure 1.1: Block diagram illustrates the primary theme of this thesis. We want to explore problems where eye tracking data can improve computer vision problems related to object extraction in images and videos. Additionally, we also want to investigate the inverse scenario of how object detection can benefit the prediction of eye tracking data in images.

Human visual attention is typically attracted towards high level semantics in images and videos. Figure 1.2 shows examples of eye tracking data obtained from a subject while performing a free viewing task in videos. As the eye tracking data is naturally biased towards objects, in this thesis we aim to improve computer vision algorithms related to object detection in images and object extraction in videos as shown in Figure 1.4. Our hybrid eye tracking/saliency based computer vision algorithms achieve sig-

nificant improvement over state-of-the-art. Eye tracking based contextual information has become practically feasible owing to recent advancements in eye tracking technology. The modern state-of-the-art eye trackers are affordable without compromising the precision of expensive laboratory eye trackers. Moreover, advancements in sensing and eye localization algorithms have enabled eye tracking acquisition without constrains on head movements effectively simulating real world viewing of scenes without affecting the experience of the viewer. Therefore, it is feasible to obtain eye tracking data from multiple subjects, especially when we are dealing with popular multimedia content. An illustration of the eye tracker used in all the experiments in this thesis is shown in Figure 1.3.

Additionally, visual saliency based contextual information only incurs a computational overhead of computing the saliency maps, however reducing the search space for object detectors can mitigate some of this additional processing. In order to improve eye tracking prediction using computer vision algorithms, we explore two forms of contextual information to improve current state-of-the-art visual attention prediction algorithms. First, we investigate the importance of scene context, utilizing object co-occurrence to improve state-of-the-art visual attention model. We also study how camera focus affects visual attention. Recently, light field cameras which capture the entire 3-D light field and camera focus can be altered post image capture, have become popular and we propose models which utilizes camera focus to identify interesting re-

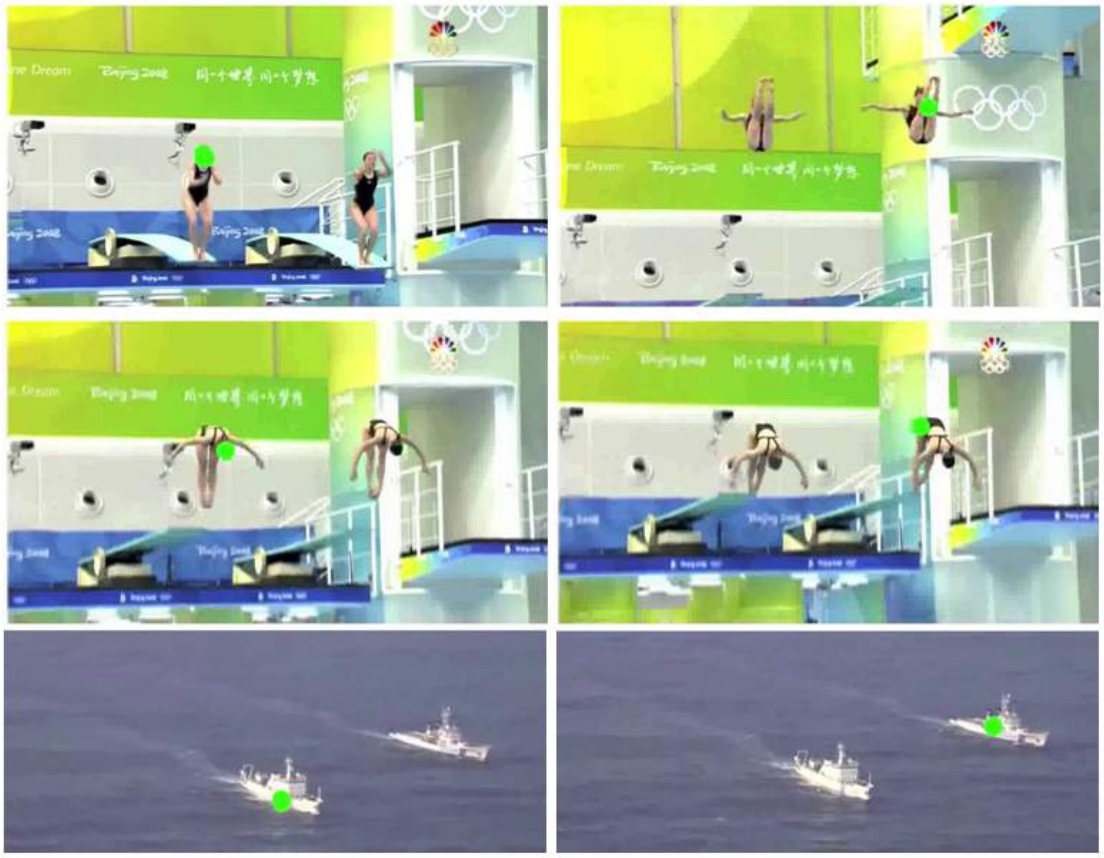


Figure 1.2: Figure overlays eye tracking data (shown as green circular dot) from a subject onto video frames from two different video sequences. We notice that eye tracking data in this free viewing task is biased towards objects in the scene and attention shifts from one object to another.

gions in images. Our method improves over state-of-the-art to predict visual attention which do not explicitly utilize this information.


Figure 1.3: This image shows the eye tracker setup used in the proposed work. Detailed description of the eye tracker setup is discussed in Chapter 2

## 1.2 Overview of proposed methods

As faces and text regions in images primarily attract visual attention, in Chapter 3, we aim to predict face and text regions by analysing eye tracking data. Additionally, we utilize this eye tracking prior to localize the search space for object categories in images. In this work we introduce a dataset with emphasis on faces and text, however containing adequate representation from non-faces and non-text as well. We collect


Figure 1.4: Lists the computer vision algorithms which we aim to improve using the hybrid eye tracking object detection algorithms. (Left) Text detection in natural scenes (Center) Dog and Cat detection (Right) Object extraction in videos

eye tracking data from multiple subjects and cluster the fixation component using mean shift clustering. We extract several inter- and intra-cluster features from these clusters. The face/text identification problem is modeled as a cluster labeling problem over a fully connected Markov Random Field. The unary and the pairwise potentials are learnt using Support Vector Machines. This prior localizes the search space for face and text regions. In the face detection problem, we focus on dog/cat faces as traditional detectors fail for these categories. Our hybrid eye tracking object detector outperforms the base algorithms in detecting faces and text.

In Chapter 4, we extend the the theme of improving object detection using eye tracking prior to videos. Specifically we target object extraction from videos and evaluate the performance gains by utilizing eye tracking priors. This problem presents interesting applications in multimedia annotation and multiple object extraction where recent advances in technology has laid the foundation to obtain eye tracking data without in-

terfering with viewer experience. This work primarily consists of two modules

- Extraction of dominant eye tracking patterns from user data
- A novel framework for multiple object extraction

First, we extract dominant consistent eye tracking patterns (visual tracks) from user data. This provides us with two forms of information, coarse localization of object position and the number of objects in the video. We utilize a two-step association process to build visual tracks. In the first step the eye tracking data is clustered using 3-D mean shift algorithm to get visual tracklets. The visual tracks are formed by associating the visual tracklets by solving a linear assignment problem. These visual tracks are utilized in the second module to extracts objects

The second module is is designed to extract multiple objects from a video based on building a mixed graph which has both directed and undirected edges. The nodes are represented by candidate bounding boxes derived from objectness, scored using a combination of objectness, optical flow and eye tracking prior. The graph consists of directed edges which connect the nodes across successive frames as well as undirected edges within (intra) a frame. The directed edges model inter-frame bounding box characteristics such as appearance, position, motion and visual track properties, enabling consistent object tracking across candidate objects. Additionally, the intra-frame costs ensure overlapping bounding boxes representing the same object is not extracted in

different paths. The overall path extraction is solved using integer programming in a mixed graph framework. The final object contours are obtained by a 3-D graph cut based video segmentation algorithm. The proposed algorithm outperforms prior dominant object extraction state-of-the-art as well as prior work on eye-tracking fixation based tracking.

In Chapter 5 , we explore the ability of saliency algorithms to improve text detection in natural scenes. Human attention is naturally biased towards text regions and we want to investigate the utility of low-level visual attention features which mimic human attention in localizing text regions in natural scenes. Additionally, text regions have characteristic visual attention properties and we aim to understand this using our algorithm. Our approach basically learns a text attention map using a Support Vector Machine from multiple visual saliency maps which prunes the search space for text detection. Our algorithm gives priority to regions where text detection typically fails. This approach improves the precision of state-of-the-art text detector.

The following Chapter 6 sheds light on improving state-of-the-art visual attention models using scene context information. We explore two forms of contextual information, object co-occurrence and camera focus to improve visual attention prediction from multiple subjects. We extract object co-occurrence maps coupled with camera focus based information to represent scene context features. A regression based attention prediction algorithm predicts the final visual attention map. The proposed algorithm

outperforms state-of-the-art saliency and visual attention algorithms as they do not directly model these contextual inter-relationships. Finally in Chapter 7, we provide the conclusions and directions for future research in saliency and eye tracking enhanced computer vision.

# Chapter 2

# Eye tracking review

"The eyes are the amulets of the mind".

*W.R. Alger*

The primary aim of this thesis is to improve computer vision algorithms using eye tracking data. This chapter reviews current work on eye tracking and describes the experimental setup used in this dissertation research.

### 2.0.1 Brief History of eye tracking

Eye tracking is the process of measuring the motion of the eye relative to the head position. The device designed to measure eye movements is an eye tracker. Eye trackers are popular devices in visual system research. Eye trackers have been extensively

used in psychology, cognitive linguistics and product design. The study of eye movements began in the 1800s. In 1879, Louie Émile Javal observed that during reading task, eye movements do not involve smooth sweeping of eyes along the text as previously understood, instead it involves a series of rapid stop and go motion patterns. This observation led to considerable research on eye movements during reading task to understand the stop duration, word positions which have consistent eye movement stops, and the duration of the stops in various scenarios. One of the early intrusive eye trackers was built by Edmund Huey which uses a contact lens with a hole for pupil. The aluminum pointer connected to the lens was used to record eye movements. His primary study was dedicated to the reading task. The first non-intrusive eye tracker was built by Guy Thomas Buswell, which recorded light rays reflected from the eyes using a film to calibrate eye movements and his research focused on reading and picture viewing.

Critical findings in eye tracking research was performed by Alfred Yarbus, where he identified that task which a subject is performing plays an important role in determining eye movement patterns. The famous image from Yarbus's work shown in Figure 2.1 elucidates the importance of top-down task influence in eye movements. He concluded that human attention is often attracted towards unusual and incomprehensible elements in images. Also, additional time spent on perception is not used to examine secondary elements, but to re-examine important details.

In the 1970s reading research took prominence again and in 1980 Just and Carpenter proposed that "There is no appreciable lag between what is fixated and what is processed" which is also called the strong eye-mind hypothesis. This hypothesis is typically assumed by eye tracking researchers and our work was well, which basically means when a subject fixates at a scene or an object, she/he also thinks about it only during the recorded fixation duration. However this assumption is questionable in situations where covert attention plays a critical role when a subject is observing which he/she is not looking at using peripheral vision. This situation however disassociates the relationship between eye movements and cognitive processing.

Due to advancement in processing power of modern computers, the 1980s also signaled the genesis of eye tracking in human computer interaction. The application domain was primarily targeted towards physically challenged users. However, recently eye tracking has been extensively used to evaluate the design of interfaces. This provides a solid platform to test the ease of use of a computer interface and helps quantify their intuitiveness. Eye tracking technology has also been a useful tool to evaluate the utility of websites to communicate information effectively. It is also becoming popular in human computer interaction (HCI) where scanpaths are utilized to build gaze-contingent displays, also known as gaze-based interfaces. Online advertising is another field where eye tracking technology will have a significant impact in the near future.

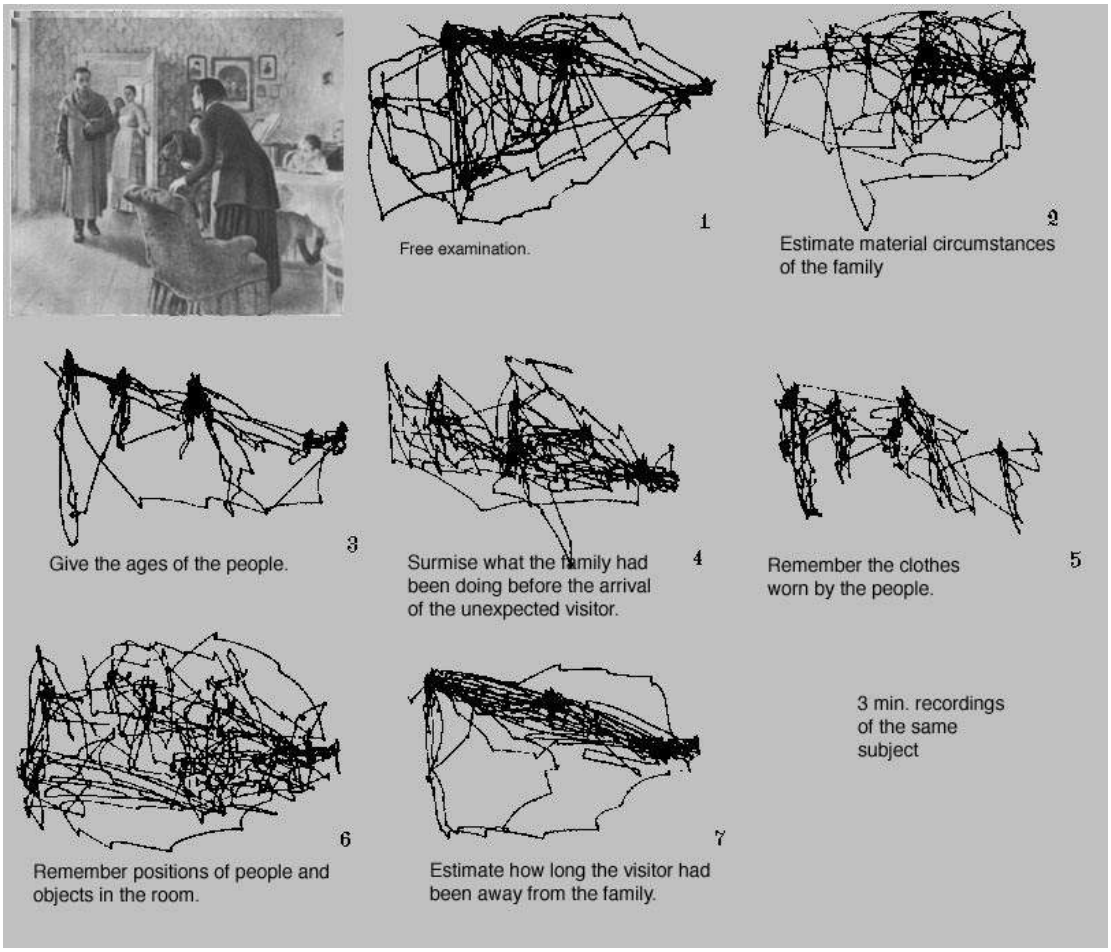



Figure 2.1: Study by Yarbus which indicates the importance of the task which the person is asked to perform on eye movements [102]

## 2.1 Eye tracking data basics

Human eye movements show considerable variation in static and dynamic scenes. In static scenes, eye movements consist of discrete jumps between information gathering stages. In Figure 2.2, which represents reading task in a static scene, we notice circles representing the information gathering stage and horizontal lines indicating shift in attention from one word to another. The circles are called fixations and the jumps from one fixation to another is called a saccade. We seldom observe smooth eye movement patterns in static scenes.

However in dynamic scenes, when a subject is observing a moving object, we notice smooth eye movement transitions, called smooth pursuit. This is another voluntary manner in which humans can shift gaze in addition to saccadic movements which is also prevalent in dynamic scenes. In this situation, smooth pursuit and fixations represent the information gathering stage from an object and the saccades denote shift in attention from one object to another in the dynamic scene.

In addition, eye tracking data also consists of micro-saccades. Micro-saccades are typically observed in prolonged visual fixations. They are small, jerk-like, involuntary eye movements, similar to miniature versions of voluntary saccades. There is still no consensus on the role of micro-saccades in visual perception, thought several theories exist. The study of micro-saccades acts as a diagnostic test for conditions such as Attention Deficit Hyperactivity Disorder (ADHD).

We note that the central one or two degrees of the visual angle (the fovea) provide the bulk of visual information; the input from larger eccentricities (the periphery) is less informative. Hence, the locations of fixations along a scanpath indicate which information regions on the stimulus are processed during an eye tracking session. On average, fixations last for around 200 ms during the reading of linguistic text, and 350 ms during the viewing of a scene. Additionally, preparing a saccade towards a new goal takes around 200 ms. Research has suggested that there is about 100-250 millisecond lag in eye movements following visual attention.

Eye tracking studies have shown that human overt visual attention, which eye trackers measure, is highly biased towards high level semantics in a scene. Additionally, among semantic objects, there is conclusive evidence that face and text categories primarily attract visual attention. Visual scanpaths are typically useful for analyzing cognitive intent, interest, and salience. Other biological factors (age, gender, race) affects the scanpath as well. However, we note that current technologies cannot predict the exact cognitive process given eye movement patterns, for example eye movements over an object may indicate several emotions, however there is no known technique to extract this information from eye movements. Therefore, additional verbal cues are sometimes obtained in eye tracking experiments.

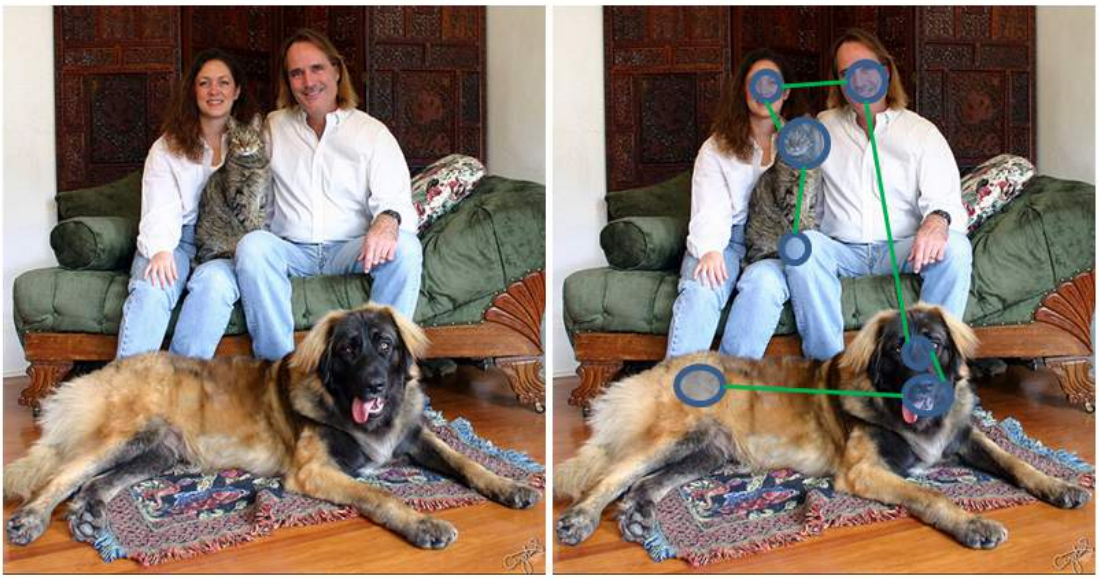


Figure 2.2: Example eye tracking data on images during a free viewing task. Eye movement data on images typically consists of alternate fixations and saccades. The fixations are represented by circles and the saccades are represented by lines. The fixations indicate information gathering stage from an image region. The saccades indicate attention shifts from one fixation to another. The entire eye movement sequence is called a visual scanpath.

## 2.2 Eye Trackers

Eye trackers are devices that measure eye movements and they fall into the following three categories

- Optical tracking
- Eye attached tracking
- Electric potential based tracking

### 2.2.1 Optical tracking

Optical tracking is the most popular form of current eye tracking technology. This is a non-contact method where infrared light reflected from the eye is captured by an optical sensor. Typically the pupil and the first corneal reflection are tracked over time. In the beginning a known calibration pattern is presented in the screen which helps localize the exact eye gaze location with good accuracy. Recently more sensitive eye trackers which monitor the location of retinal blood vessels are also being developed. Optical tracking technology has required head position to be stable when tracking the eye movements. Recent developments in calibration and head tracking has enabled free head movement while localizing eye position. This is a significant step which enables collection of eye tracking data over long duration without discomfort to the subject. These remote eye trackers automatically track the head as well as the eye movements

are becoming commercially available . In addition manufacturing costs of the eye tracking devices have drastically reduced in the past few years thereby improving the utility of these eye trackers in the commercial sector and going beyond being expensive laboratory equipment. Optical eye tracking technology has been used in all the experiments conducted in this dissertation work.

Head mounted eye trackers and chin rest based eye trackers which require the head to be stable have been popular in the past decade. Early eye trackers used a sampling rate of at least 30 Hz. Today many video-based optical eye trackers run at even 1000/3000 Hz, which is needed in order to capture the details of the very rapid eye movement in neurological studies.

### 2.2.2 Electric potential based tracking

Eyes act as a dipole with the positive pole in the cornea and the negative pole at the retina. This produces a steady electric field which is measured by placing electrodes around the eyes. Two pairs of contact poles called an electrooculogram (EOG) are placed around the eye which measures this electric signal. The dipole orientation changes when the eyes move from the center position to the periphery. As the retina and the cornea approaches opposite electrodes, there is a change in the measured EOG potential signal. These changes can be related to eye movements. Typically horizontal and vertical potentials are measured by separate electrodes. A third EOG component is

the radial EOG channel, which is the average of the EOG channels referenced to some posterior scalp electrode. This radial EOG channel is sensitive to the saccadic spike potentials from the extra-ocular muscles at the onset of saccades, and allows reliable detection of even miniature saccades. An example EOG based eye tracking setup is shown in Figure 2.3.

However the primary limitation of EOG is that it cannot accurately predict where exactly a person is looking at due to noise in electric potential measurements. But, it can accurately measure saccadic eye movements associated with gaze shifts and blinks. It consumes low power, is available as a wearable system and is robust to lighting conditions. In addition, sleep activity can be monitored as it does not require the eyes to be open while recording the eye activity.

### 2.2.3 Eye attached tracking

In this technique an attachment to the eye is used and the movement of the attachment is measured which indicates eye movements. Typically specialized contact lenses which have embedded magnetic field sensor or mirror is used as attachment. However, eye attached tracking assumes that the contact lens does not slip during eye rotation.

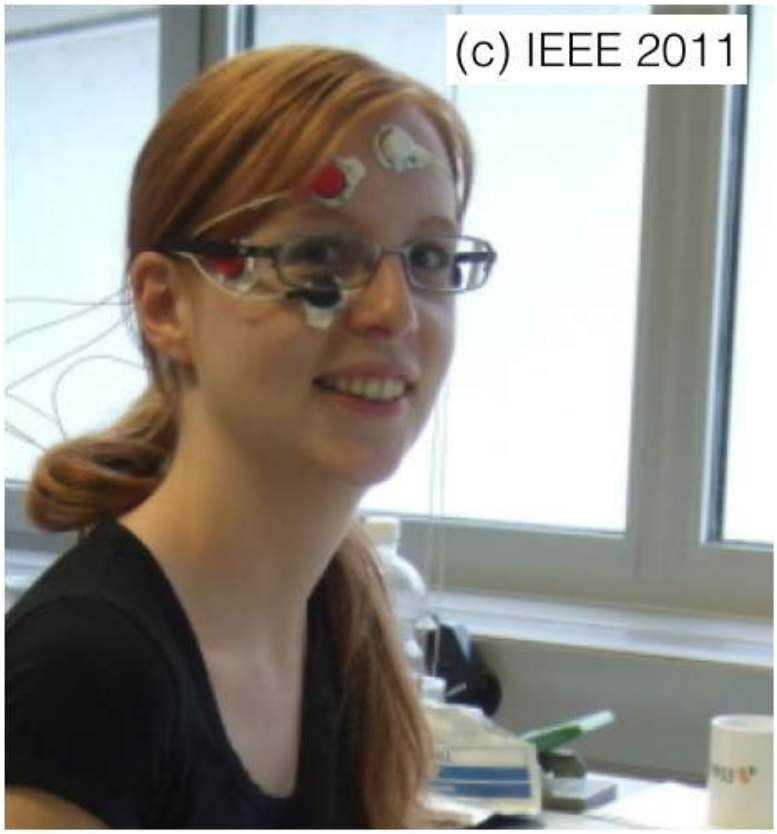



Figure 2.3: Example of electrooculography based eye tracker [16].

## 2.3 Eye tracking setup

In our experiments we use the video based eye tracker Eyelink 1000 as it can accurately track where subjects look in a variety of setups. Specifically we utilize the tower mounted setup with a chin-rest as it ensures comfortable viewing in a short period of time coupled with high eye tracking accuracy. The eye tracking setup consists of a host computer and a display computer. The eye tracker host computer processes the eye tracking data and automatically computes fixations and saccades in real time at a high sampling rate of about 1000 samples per second. The host computer is a part of the Eyelink eye tracker.

The display computer is a device where the visual stimuli is presented to the viewer, which is typically a computer used by the programmer to design the eye tracking experiments. The host and a display computers are connected using a duplex cable which can transfer data to and from the host and display computers. The eye tracker also consists of an infra-red camera which collects an infra-red video which is processed by the host computer. The host computer detects the pupil and the first corneal reflection to decipher where a subject is looking in the display screen. To enable this, calibration is performed every time a subject places his/her head on the chin-rest where a series of 9 points is shown in the display screen and the pupil and the first corneal reflection positions are monitored.

The eye tracking experiments are built using an experiment builder software pro-

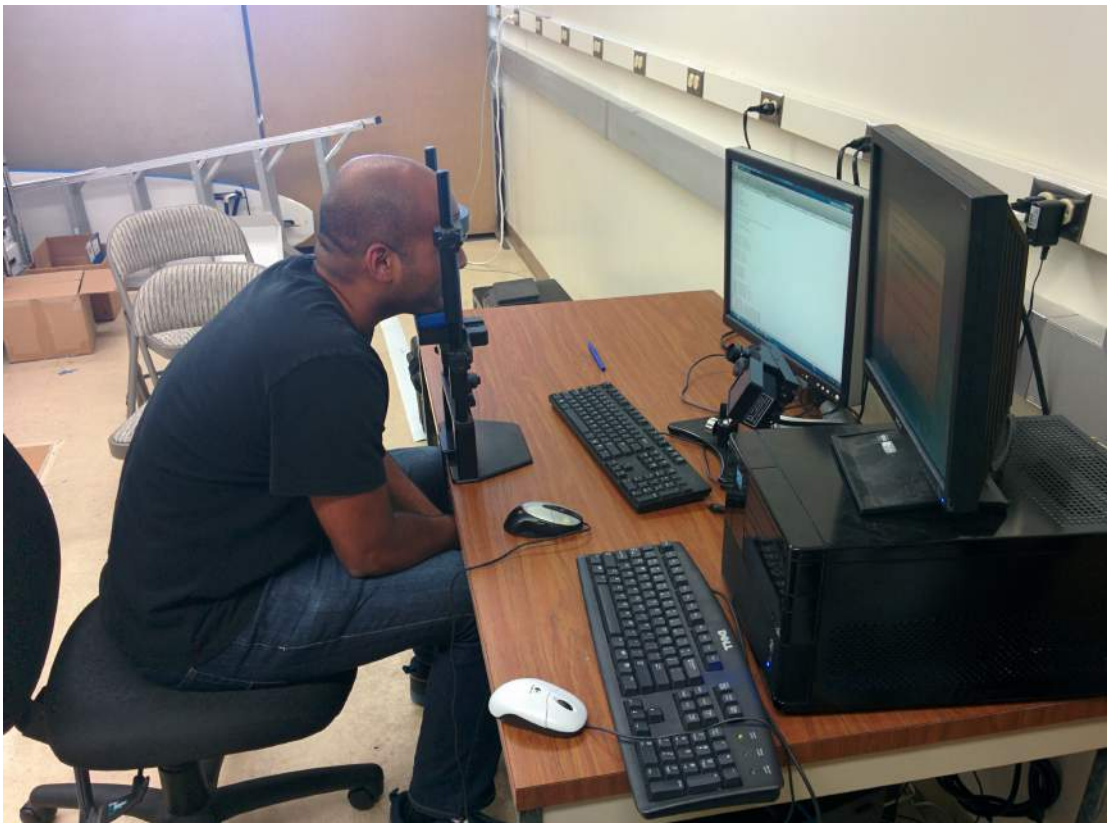


Figure 2.4: Illustrates the eye tracker setup using Eyelink 1000 eye tracking device

vided by SR-Research, the company which manufactures Eyelink. The experiment builder software provides a few basic block based setups for collecting eye tracking data and it acts a good guiding tool to learn how one can design eye tracking experiments. In addition, to design complex experiments where the data presentation might be adaptive, psychtoolbox is recommended instead. In addition, SR-Research also provides a data viewer to visualize the eye tracking data overlaid on the stimulus post data collection from a subject.

We now outline briefly the work-flow in an eye tracking experiment. After identifying the hypothesis which we want to test, we collect a visual stimuli database, following which the experiment is designed using the experiment builder software. As fatigue plays an important role in the chin-rest eye tracker (head is constrained in a specific position), one must ensure that every data collection session is divided into smaller sub-sessions and the subjects gets sufficient rest before continuing the experiment. The eyetracking data for each subject is stored in EDF (Eyelink Data File) format. This is a highly compressed binary format, intended for use with SR Research Eyelink viewers and applications. The EDF file can be converted to a text based ASC file by a translator program known as EDF2ASC. The ASC files may be viewed with any text editor. Further, we wrote Perl scripts to process the ASC files line by line to extract useful information which we required for further analysis.

## 2.4 Eye tracking in proposed research

There has been limited research to tap into the potential of eye trackers in computer vision. However, several eye tracking based research studies have concluded that high level semantic categories primarily attract visual attention and therefore eye tracking naturally can provide weak supervision in several computer vision problems related to object search. In Chapters 2 and 3 of this thesis we design algorithms which

can outperform state-of-the-art object search algorithms in images and videos utilizing eye tracking information obtained in a free viewing scenario. We hope the proposed paradigm eventually bridges the semantic gap between computer vision algorithms related to object search and human performance.

# Chapter 3

# Eye tracking enhanced object detection in images

> "We see the world, not as it is, but as we are - or, as we are conditioned to see it".
>
> *Stephen R. Covey*

Eye movement studies have confirmed that overt attention is highly biased towards faces and text regions in images. In this work we explore a novel problem of predicting face and text regions in images using eye tracking data from multiple subjects. The problem is challenging as we aim to predict the semantics only from eye tracking data without utilizing any image information. The proposed algorithm spatially clusters eye

tracking data obtained in an image into different coherent groups and subsequently models the likelihood of the clusters containing faces and text using a fully connected Markov Random Field (MRF). Given the eye tracking data from a test image, the learnt MRF predicts potential face/head (humans, dogs and cats) and text locations reliably. Furthermore, the approach can be used to select regions of interest for further analysis by object detectors for faces and text. The hybrid eye position/object detector approach achieves better detection performance and reduced computation time compared to using only the object detection algorithm. We also present a new eye tracking dataset on 300 images selected from ICDAR, Street-view, Flickr and Oxford-IIIT Pet Dataset from 15 subjects.

This Chapter is organized as follows. In Section 3.1 we introduce the problem and relevant related work. The faces and text eye tracking dataset is presented in Section 3.2. The proposed approach to classify eye tracking data into face and text regions is described in Section 3.3 followed by the experimental results for eye tracking based localization in 3.4. The applications to improve state-of-the-art object detection algorithms is illustrated in 3.5. Finally, the summary of the work and future research directions are discussed in Section 3.6.

## 3.1 Introduction

Wearable eye tracking devices are becoming popular [15, 14] and will soon be mainstream. They provide a platform to collect eye tracking data in a non-intrusive way when people observe multimedia content, such as web browsing. This additional information from multiple subjects can potentially be useful for challenging large scale multimedia annotation problems. Towards this, we propose a technique to obtain image-level scene semantic priors from eye tracking data, which will reduce the search space for multimedia annotation tasks.

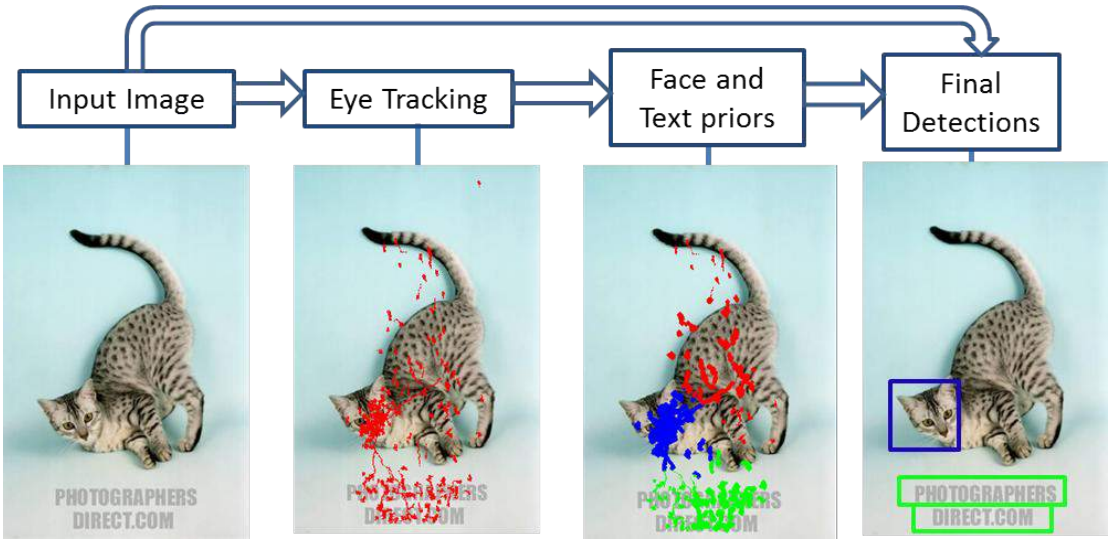



Figure 3.1: Left to right: 1. Input image. 2. Eye Tracking fixation samples from multiple subjects overlaid on the image 3. The eye tracking regions identified by the proposed algorithm as faces (blue) and text (green) 4. The final detection outputs of face and text detector focusing on the priors provided by eye tracking.

It is known that human visual attention, irrespective of top-down task, is biased to-

wards faces and text [17]. The first step towards obtaining scene semantic prior from eye tracking information alone is to build models that predict face and text regions in images, which is the primary focus of this work. This information is useful to improve the speed and precision of state-of-the-art detectors for challenging categories such as text, cats and dogs. We note that the performance of state-of-the-art cat and dog detectors [74] in turn depends on head (face) detection algorithm which can be enhanced using eye movement information.

**Related Work**

Humans are able to swiftly process a rich stream of visual data and extract informative regions suitable for high level cognitive tasks. Therefore, there has been significant amount of research on human inspired visual attention models [51, 39, 53, 56]. These approaches typically predict the attention in different regions of an image given low-level saliency maps and high-level image semantics. In contrast, the proposed problem in spirit models the converse situation of predicting image semantics from eye movement data.

There have been some recent efforts which model top-down semantics by simultaneously utilizing both image and eye movement information. In this regard, Subramanian et al. [86] extract high-level information from images and verbal cues, (faces, face parts and person) and model their interrelationships using eye movement fixations and saccades across these detections. Mishra et al. [68] propose an active segmentation

algorithm motivated by finding an enclosing contour around different fixations. The proposed approach distinguishes itself as it aims to speed up algorithms for high-level semantics from eye movement data alone. Bulling et al. [16] propose an activity classification method in office environments (copying text, reading, browsing web, taking notes, watching video) using eye movement data collected using electrooculography. As most of these activities follow a standard repetitive pattern, the method in [16] predicts the activities reliably for each person individually. However, due to variability in the manner in which different people view images, our approach differs from [16] and we require data from multiple observers to predict image semantics reliably. Cerf et al. [18] provide an algorithm to decode the observed image using eye movement scanpath data. However, their approach models the problem by proposing a metric between multiple saliency maps obtained from the image and the scanpath data. The saliency map generation problem again requires processing the entire image and is inherently different from the proposed approach. We make three contributions in this work

a. We propose an algorithm to localize face and text regions in images using eye tracking data alone. The algorithm basically clusters the eye tracking data into meaningful regions using mean-shift clustering. Following which various intra- and inter-cluster fixation and saccade statistics are computed on these clusters. The final cluster labels are inferred using a fully connected MRF, by learning the unary and interaction potentials for faces and text from these statistics.

b. We demonstrate the ability of these face and text priors to improve the speed and precision of state-of-the-art text [32] and cat and dog detection [74] algorithms.

c. We also present a new eye tracking dataset, collected on images from various text, dogs and cats datasets. The dataset consists of 300 images from 15 subjects.

Figure 3.1 outlines the pipeline of the proposed approach.

## 3.2 Faces and Text Eye Tracking Database

We collected an eye tracking dataset, with primary focus on faces (humans, dogs and cats) and text using Eyelink 1000 eye tracking device. The image dataset consists of 300 images collected from ICDAR datasets (text) [61], Street view dataset (text) [96] and Oxford-IIIT Pet dataset (dogs and cats) [75] and flickr images [53]. The text images are gathered from two different datasets to ensure considerable variability in scene context. The flickr images provide sufficient representation for images without text or faces (including dogs and cats) in both indoor and outdoor scenes. The overall image dataset consists of 61 dogs, 61 cats, 35 human faces, 246 text lines and 63 images without any text or faces. Figure 3.2 highlights examples for images from different categories from the dataset. The images are of dimension $1024 \times 768$ and were viewed by 15 subjects (between ages 21 and 35). The viewers sat 3 feet away from a 27 inch screen and each image was shown for 4 seconds followed by 1 second viewing

a gray screen. The subjects were informed that it was a free viewing experiment and instructed to observe regions in images that gather their interest without a priori bias. Also, eye tracking calibration was performed every 50 images and the entire data was collected in two sessions (150 images each). This dataset can be downloaded from http://vision.ece.ucsb.edu


Figure 3.2: Examples of images from our dataset consisting of text, human faces, dogs, cats and other background objects

Humans eye movement scanpaths typically consists of alternating fixations and saccades. Fixations represent information gathering sequences around an interest region and saccades indicate transitions between fixations. The eye tracking host computer samples the gaze information at 1000 Hz and automatically detects fixations and saccades in the data. Therefore, we have around 4000 samples per subject for every image. The fixation samples typically account for 80% of the entire data. In our analysis we

only use the fixation samples and henceforth refer to these fixation samples as the eye tracking samples. The eye tracking device also clusters the fixation samples and identifies fixation and saccade points. We refer to these points as fixations and saccades hereafter. The average number of fixations and saccades per image across subjects can vary from 8 to 19. In our experiments, the first fixation and saccade was removed to avoid the initial eye position bias due to the transition gray slide in the experimental setup.

Face Regions: The dataset consists of faces of multiple sizes, varying from about 40×40 to 360×360 pixels. In small face images, subjects look at the face as a whole. On the other hand, in larger faces there are several saccades across eyes, nose and mouth regions. As expected, face regions consistently attract attention from viewers. In addition we notice that the initial saccades are invariably directed towards face regions across subjects. In images consisting of multiple faces, rapid scanpaths moving across different faces is a common phenomenon. Figure 3.3 illustrates examples featuring some of these effects.

Text Regions: Text regions are present in various styles, fonts, sizes, shapes, lighting conditions and with occlusions from foreground objects in our image dataset. In text regions consisting of a single word, the subjects typically fixate around the center of the word and the different fixations take a nearly elliptical shape. In multiple words, we observe saccadic scanpaths from one word to another as subjects typically read the

different words sequentially. Figure 3.3 illustrates some example text regions in our image dataset.

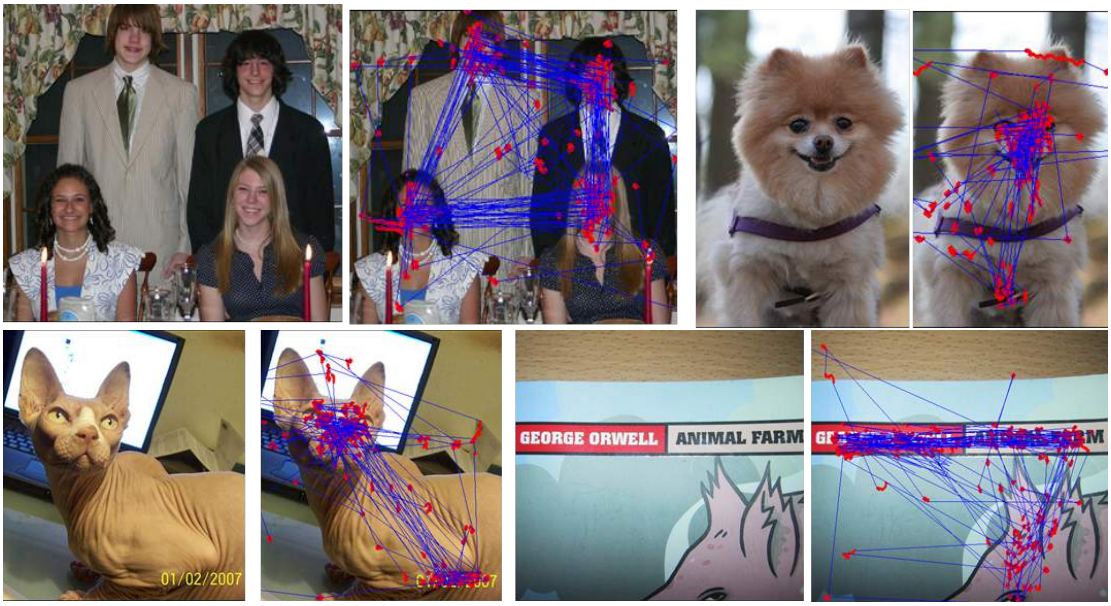



Figure 3.3: Shows example of faces and text in two scenarios each. The fixations are marked as red points and saccades as blue lines. Multiple faces in the image where we consistently observe inter-face saccades (top left).A large single face where several saccades are observed in the eyes, nose vicinity (top right). Text with four words where a dense saccadic presence is observed between words (bottom left).A clip from one of the images showing a single word, whose cluster takes a nearly elliptical shape (bottom right).

## 3.3 Faces and Text Localization from Eye Tracking Data

The aim is to identify face and text regions in images by analyzing eye tracking information from multiple subjects, without utilizing any image features. Eye movements are organized into fixations and saccades. The information gathering phase is rep-

resented by the fixations, which typically group around different semantic/interesting regions as shown in Figure 3.3. Therefore, we first cluster all the fixation regions using the mean-shift clustering technique [25]. We chose mean-shift clustering as it does not require the number of clusters and is fairly robust to multiple initializations for the selected bandwidth (50 pixels). The text and face region detection problem is mapped to a cluster labeling problem. Therefore, we compute inter-cluster and intra-cluster statistics and model the labeling problem using a fully connected Markov Random Field (MRF).

Let the $i^{th}$cluster in an image be denoted by $\mathcal{C}_i$. The 2D eye tracking samples (fixation samples) within the cluster are represented by $E_i$. The fixations (fixation points) and saccades in the entire image are denoted by $\mathcal{F}$ and $\mathcal{S}$ respectively. The fixations belonging to the $i^{th}$ cluster are denoted by $F_i$ and the saccades originating from $i^{th}$ and terminating in the $j^{th}$ by $S_{i,j}$. Finally, the fixations provided by every individual person $k$ in cluster $i$ is augmented giving $F_i^k$ and the corresponding times (0-4 seconds) representing the beginning of the fixations in cluster $i$ is given by $T_i^k$. The following features are used to represent inter-cluster and intra-cluster properties.

### 3.3.1 Intra-cluster features

*a.* Number of fixations and eye tracking samples: $|F_i|$, $|E_i|$

*b.* Standard deviation of each dimension of the eye tracking samples $E_i$

*c.* Shape and orientation of the cluster by ellipse approximation. Let $\lambda_1$,$\lambda_2$ and $v_1$,$v_2$ denote the two eigenvalues and eigenvectors respectively of the cluster such that $\lambda_1 > \lambda_2$. Shape of the cluster is encoded by $\frac{\lambda_2}{\lambda_1}$. The orientation is expressed as $|\angle v_1|$

*d.* The ratio of the eye tracking sample density in the cluster compared to its background. Let cluster $\mathcal{C}_i$ be approximated by the minimum rectangle $R_i$ containing all the cluster points. The rectangular region centered around $R_i$ which is twice its width and length is defined as $D_i$. Hence, the background region, $B_i$, around $R_i$ is expressed as $D_i \setminus R_i$. The final feature is computed as $\frac{|\{E_i \in B_i\}|}{|\{E_i \in R_i\}|}$

*e.* Number of incoming, outgoing and within-cluster saccades, represented by $\sum_{\forall j \neq i} |S_{j,i}|$, $\sum_{\forall j \neq i} |S_{i,j}|$ and $|S_{i,i}|$ respectively

*f.* The number of incoming, outgoing and within-cluster saccades, (from *e*) where the saccade angle to the X-axis is less than 30 degrees (almost horizontal)

*g.* The percentage of incoming, outgoing and within-cluster saccades (from *e*) which are almost horizontal

*h.* Median of the time of first visit to the cluster across multiple subjects: $\text{median}_k \left(\min_i \left(T_i^k\right)\right)$

*i.* Median of the number of times each person visits a cluster: $\text{median}_k(\left|F_i^k\right|)$

In total we have 18 intra-cluster features representing each cluster's intrinsic properties.

These features essentially aim to capture the eye movement attributes typical of face and text regions described in Section 3.3. The features indexed *a,b,c,d* and *e* in the above list are important basic features where text and face regions exhibit characteristic responses. Features *f* and *g* are more characteristic of text regions with multiple words as nearly horizontal inter-word saccades are prevalent. Finally, features *h* and *i* are more relevant to face regions which typically immediately attract viewer attention. In addition subjects also tend to visit the face multiple times after fixating at other regions in the image, which is captured by feature *i* in the aforementioned list.

### 3.3.2 Inter-cluster features

In addition to intra-cluster features, pairwise inter-cluster features also provide useful information to identify face and text regions. In the presence of multiple faces, subjects indicate saccadic activity across the different faces. Moreover, in text images with multiple words, inter-word saccadic activity is quite common. Therefore, the following saccade centric features are computed across clusters.

*1.* Number of saccades from the $i^{th}$ to $j^{th}$ cluster, $|S_{i,j}|$ and vice versa

*2.* Number of almost horizontal saccades (where the saccade angle to the X-axis is less than 30 degrees) from cluster $i$ to $j$ and vice versa

*3.* Percentage of almost horizontal saccades from cluster $i$ to $j$ and vice versa

4. The number of saccades, horizontal saccades and percentage of horizontal saccades from the left cluster to the right cluster

5. Distance between the clusters

In total, we have 13 inter-cluster features to represent saccadic properties across multiple clusters. Specifically, the inter-cluster features *1,2* and *3* from the list above are useful indicators of face-face and text-text regions. Also, feature *4* is targeted to capture text regions as subjects typically read text from left to right.

**Data**: Input Images $\{\mathcal{I}^i\}$, Eye Tracking Samples $\{\mathcal{E}^i\}$, Fixations $\{\mathcal{F}^i\}$, Saccades $\{\mathcal{S}^i\}$ ground truth labels for faces and text $\{\mathcal{L}^i\}$, $i \in [1...N]$

**Result**: Face Cluster IDs$^i$, Text Cluster IDs$^i$,$i \in [1...N]$

**Notation :** Superscript - image number. Subscripts - cluster IDs

Precomputing Cluster Features:

**for** $i = 1 \rightarrow N$ **do**

  $\mathcal{C}^i$ = Mean Shift Clustering($\mathcal{E}^i$);

  **for** $j = 1 \rightarrow |\mathcal{C}^i|$ **do**

    $\mathcal{F}intra^i_j$ = Intra-cluster-features($\mathcal{C}^i_j, \mathcal{F}^i_j, \mathcal{S}^i_j$) ;

    $\mathcal{C}lab^i_j$ = Cluster-labels($\mathcal{L}^i_j, \mathcal{C}^i_j$);

    **for** $k$=$j+1 \rightarrow |\mathcal{C}^i|$ **do**
      $\mathcal{F}inter^i_{jk}$ = Inter-cluster-features($\mathcal{C}^i_j, \mathcal{F}^i_j, \mathcal{S}^i_j, \mathcal{C}^i_k, \mathcal{F}^i_k, \mathcal{S}^i_k$)

    **end**

  **end**

**end**

Learning to classify Clusters into Face and Text regions:

**for** $i = 1 \rightarrow N$ **do**

  TestIndex = $i$; TrainIndex = $\{1, 2, ..., N\} \setminus \{i\}$;

  [Unary Potentials Face, Unary Potentials Text] =

  QDA($\mathcal{F}intra^{\text{TestIndex}}, \mathcal{F}intra^{\text{TrainIndex}}, \mathcal{C}lab^{\text{TrainIndex}}$);

  [Pairwise Potentials Face, Pairwise Potentials Text]=

  QDA($\mathcal{F}inter^{\text{TestIndex}}, \mathcal{F}inter^{\text{TrainIndex}}, \mathcal{C}lab^{\text{TrainIndex}}$);

  Face Cluster IDs$^i$ = MRF$_{\text{face}}$(Unary Pot. Face, Pairwise Pot. Face);

  Text Cluster IDs$^i$ = MRF$_{\text{text}}$(Unary Pot. Text, Pairwise Pot. Text);

**end**

**Algorithm 1:** Proposed method to detect face and text regions by analyzing eye tracking samples.

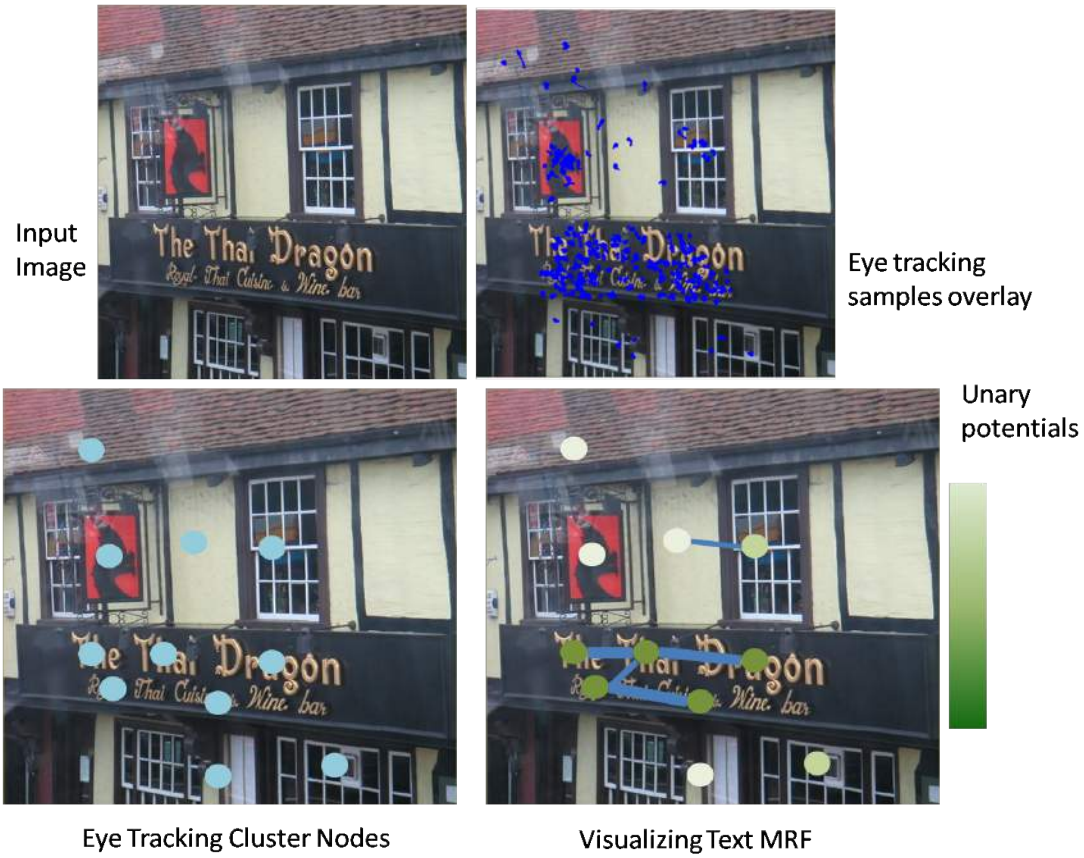



Figure 3.4: Visualizing the text MRF potentials. 1. (Top left) Input image. 2. (Top right) Eye tracking samples overlay 3. (Bottom Left) Clustered eye tracking fixation locations from multiple subjects overlaid on the image 3. (Bottom Right) Visualizing the unary and interaction potentials of the clusters for the text MRF. The unary is color coded as green, the bright values indicating high unary potentials of a cluster belonging to text class. The interaction is marked by the blue lines between clusters, whose thickness is indicative of text-text interaction magnitude.

### 3.3.3 Learning Face and Text regions

Utilizing the features in Section 3.3.1 and Section 3.3.2, we propose a probabilistic model based technique to label the clusters provided by mean-shift algorithm [25] on the eye tracking samples. The intra- and inter-cluster features are naturally modeled as a MAP inference problem using a MRF. The different clusters represent the nodes of the graph. The intra-cluster and inter-cluster features facilitate the learning of unary and pairwise potentials respectively. In addition, we utilize a fully connected graph to ensure long range interactions. Let the posterior probabilities of a quadratic discriminant analysis (QDA) classifier on intra-cluster features be denoted by $p$, the unaries are calculated as $-log(p)$. Similarly the pairwise potential is obtained as $-log(q)$, where $q$ is the posterior learnt from the inter-cluster features using QDA. The problem of inferring the labels $y_i$ of $\mathcal{C}_i$ is modeled by an MRF with energy

$$E = \sum_{i \in \mathcal{C}} V_i(y_i) + \sum_{i,j \in \mathcal{C}, i \neq j} V_{ij}(y_i, y_j) \tag{3.1}$$

where $V_i$ denotes the unary potential of cluster $i$ and $V_{ij}$ denotes the scaled pairwise potential between clusters $i$ and $j$ with a scaling factor $\lambda$. In order to allow overlapping text and face regions (in watermarked images), cope with limited availability of data with face-text co-occurrence, and speed up inference, we resort to separately tackle the face,non-face and text,non-text problems using two distinct MRFs. Finally, as we are dealing with a binary inference problem on limited number of clusters ($< 20$), we

utilize fast exact inference by pre-computing all the possibilities for different number of nodes. Figure 3.5 indicates how the potentials are learnt from the inter- and intra-cluster features. Also, Algorithm 1 enumerates the steps involved in learning face and text clusters from eye tracking data.

The algorithm formally describes that the eye tracking data is clustered using mean-shift clustering algorithm. For each cluster, intra-cluster features ($\mathcal{F}intra$) are computed according to Section 3.3.1 and for each pair of clusters, inter-cluster features ($\mathcal{F}inter$) are computed according to Section 3.3.2. The intra-cluster features help learning the unary potentials and the inter-cluster features facilitate learning the pair-wise potentials respectively using Quadratic Discriminant Analysis (QDA) classifiers. The parameters of the classifiers are learnt on a training set. Finally, the face, text and background labels of the clusters, from eye tracking data from a test image, are inferred over the MRF ($\text{MRF}_{\text{face}}$, $\text{MRF}_{\text{face}}$) by enumerating all possibilities.

## 3.4 Performance of Face and Text Localization

In this section we analyze the performance of the cluster-level classification of faces and text regions in images. To enable this, we require cluster labels from ground truth bounding box annotations. The cluster labels are defined as the the label of the class (face, text and background) which has the most representation among the cluster sam-

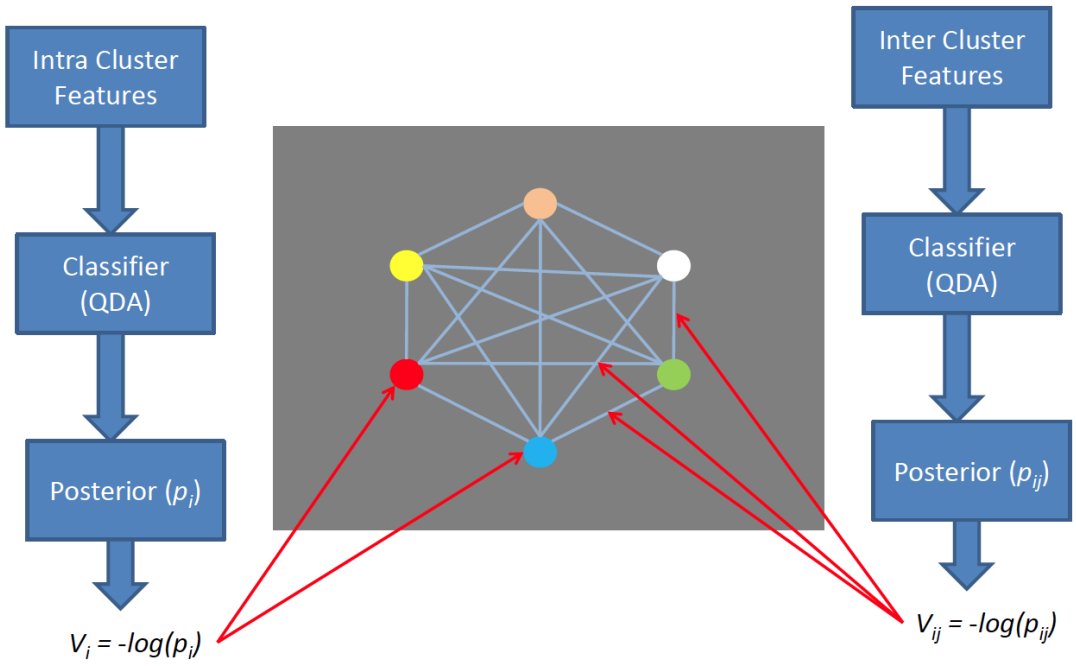



Figure 3.5: Indicates the process to calculate the unary and the pairwise potentials. The unary potentials are computed as $-log$ of posterior probabilities obtained from a quadratic discriminant analysis classifier from intra-cluster features. In similar vein, the pairwise potentials are obtained from inter-cluster features

ples. Figure 3.6 shows an example of cluster labels obtained from ground truth boxes. For this experiment we fix the bandwidth of both the face and text MRFs to 50. The parameter $\lambda$ which weighs the interaction relative to the unary potentials is fixed as $\frac{1}{|\mathcal{C}^i|}$ (to roughly give equal weights to unary and pairwise potentials), where $\mathcal{C}^i$ is the set of all clusters in the $i^{th}$ image. In addition, clusters which have less than $1\%$ of the total number of eye tracking samples are automatically marked as background to avoid trivial cases. The total number of clusters range from 3 in low entropy images to 17 in high entropy images.

Figure 3.6: Left: Input image with the ground truth for face (blue) and text (green). Center: Clustered eye tracking data overlay on input image. Right: Face (blue) and text (green) cluster labels propagated from ground truth.

The performance of the cluster detection problem is evaluated using a precision-recall approach for face and text detection. Precision and recall are defined as follows

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \tag{3.2}$$

where $TP, FP$ and $TN$ denote true positive, false positive and true negative clusters respectively in the detection problem. Finally, to get a single measure of the performance, F-measure is defined as the harmonic mean between precision and recall. In order to utilize these cluster labels to enhance text and cat and dog detection algorithms, we require high recall under reasonable precision. This ensures most of the regions containing faces and text are presented to the detector, which will enhance the overall performance.

The performance of the face and text detector MRFs are shown in Table 3.1. The results are evaluated at two levels, cluster and image. The image level metric evaluates the presence of at least one face/text region in an image. The cluster level metric

(a) Single Face Cluster Detection (b) Multiple Face Clusters Detection

Figure 3.7: Examples of good face detections from the proposed algorithm. Red fixation points correspond to face and blue corresponds to background. (a) In the presence of salient distracting object (shoe) the face (cat) is reliably detected. (b) We notice that even in challenging scenarios where multiple faces are present, the proposed approach detects reliably.


(a) Multiple Words Text Clusters Detection (b) Single Word Text Cluster Detection

Figure 3.8: Examples of good text detections from the proposed algorithm. Red fixation points correspond to text and blue corresponds to background. (a) Text line is reliably detected even in the presence of several other fixations near the region of interest. (b) Text is detected correctly in the presence of more salient object (person face).

evaluates the presence of face/text in every cluster. We notice that the recall is high for both face and text detection sections. However, the precision of the face detector is also quite high (both cluster and image level), indicating that the proposed algorithm is confident about the regions which it detects as a face. Figure 3.7 shows some example images where the proposed approach localizes faces well. Similarly Figure 3.8 highlights some text cluster detection examples. Figure 3.9 also highlights a few failure

(a) Missed Face Detection (b) Missed Text Word

Figure 3.9: Example scenario where the proposed approach fails to detect face (left) and a text word (right). The eye tracking samples detected as face in (a) and text in (b) are shown in red and the samples detected as background (both (a) and (b)) are indicated in blue.

cases where both the face and text localization fails. The face detector fails as many subjects do not concentrate on the face in the corner of the image. In addition the text cluster detection fails as the allocated time (4 seconds) was insufficient to scan the entire text content.

| | Precision | Recall | F-Measure |
|---|---|---|---|
| Face Detection Cluster | 0.671 | **0.954** | 0.788 |
| Text Detection Cluster | 0.748 | **0.942** | 0.834 |
| Face Detection Image | 0.755 | **0.989** | 0.856 |
| Text Detection Image | 0.610 | **0.923** | 0.735 |

Table 3.1: Indicates performance of cluster and image level face and text detection from the eye tracking samples.We notice that the recall (marked in bold) is high suggesting that the proposed approach seldom misses face and text detections in images. This is achieved at a sufficiently good precision ensuring that this method can be valuable to localize ROI to reduce the search space for computationally expensive face and text detectors.

## 3.5 Applications

There have been several efforts to model context [90, 7, 29, 44, 27] in single and multi-class object detection problems. The proposed faces and text eye tracking priors can be an extremely useful alternate source of context to improve detection. Therefore, we investigate the utility of these priors for text detection in natural scenes as well as cat and dog detection in images which are challenging problems.

### 3.5.1 Detecting Cats and Dogs

Detecting cats and dogs in images is a difficult task as they have high variability in appearance and pose coupled with occlusions. However, in these problems, the animal face/head is the most distinctive part and the state-of-the-art cat and dog detection algorithm proposed by Parkhi et al. in [74] makes use of this information. The final detection algorithm consists of two steps, the head/face detection and segmenting the cat/dog body by learning features from the face. The head detection used deformable parts model [35] and the body segmentation utilized iterative graph cuts [79, 11] by learning foreground and background properties . For a detailed review of the approach we refer the reader to [74].

The proposed eye tracking based face detection prior can significantly reduce the search space for cat and dog faces/heads in images. As human fixations are typically focused towards the eyes and nose of the animals, we construct a bounding box around

the face clusters to localize the cat head. When the cluster is approximated by a rectangular bounding box $R$ with width $w$ and length $l$ containing all the eye tracking samples, an outer bounding box $B$ centered around $R$ of size $2.7l \times 2.2w$ always contained the entire face within the box. Even under this conservative approximation, the search space for cat/dog faces is reduced to $15.3\%$ of the entire dataset (image area) using the proposed eye tracking based face detection model.

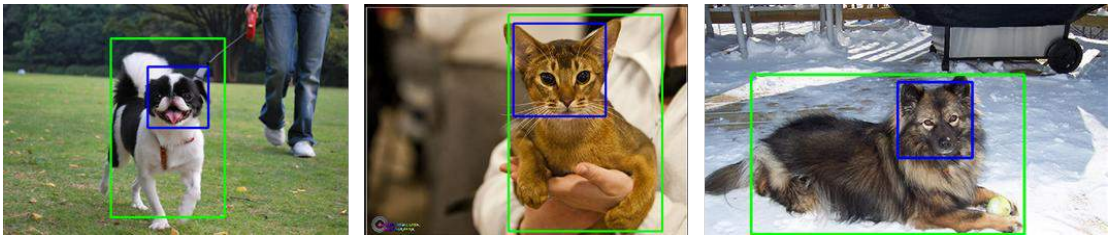


Figure 3.10: Example cat and dog face (blue box) and body (green box) detections from the proposed algorithm.

Figure 3.11 shows the Average Precision curves using multiple detection thresholds for the head detection for both cats and dogs. We notice that the head detection performed only in the rectangular regions $B$ is consistently higher than baseline (in the entire image). Especially in high recall scenarios (low detection threshold), the average precision of the proposed approach is significantly greater than the baseline approach [74]. In the whole body detection problem as well, the proposed approach outperforms the baseline approach over a larger detection threshold range. In addition, *the cat and dog head detection algorithms are 4.8 and 5.7 times faster respectively* as they operate

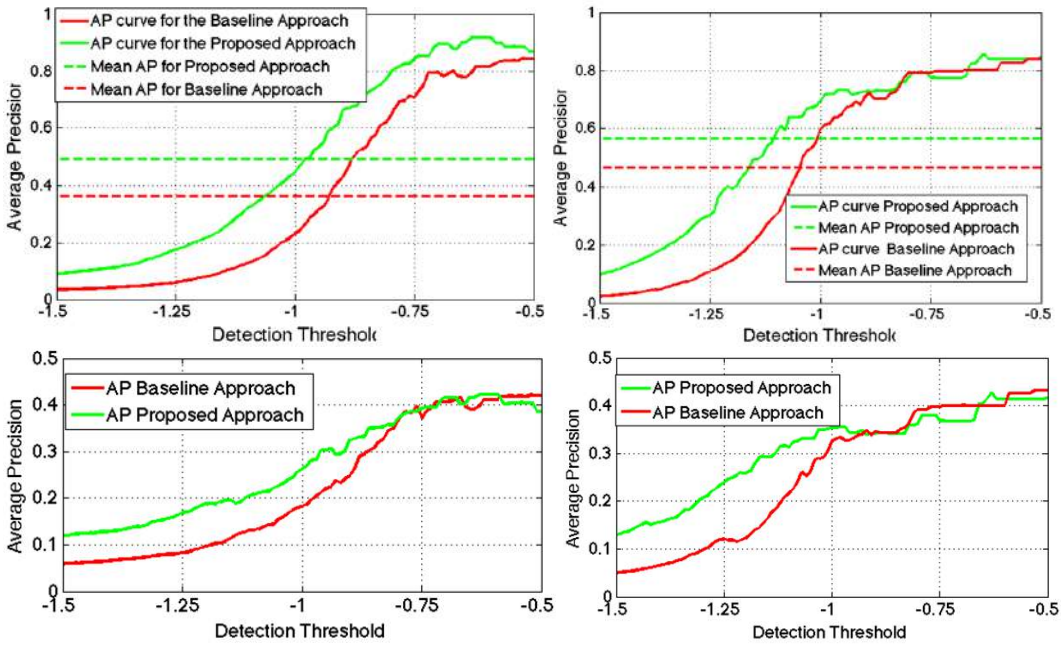



Figure 3.11: Plotting Average Precision (AP) of Cat head (top left) Dog head (top right), Cat Body (bottom left) and Dog Body (bottom right). The proposed (green) and baseline (red) curves are plotted against the detector threshold of deformable parts model. The maximum AP of baseline and proposed algorithm is comparable in all cases, however, the AP of the proposed approach is higher than baseline in high recall scenarios (low detector threshold) for both the head and body detectors of cats and dogs. Therefore, on an average the proposed approach is more stable over the detector threshold parameter than the baseline.

in the reduced search space. Therefore, we achieve dual benefits of better detection performance with considerable speed-up for dog and cat detection problems. We note that the time of the proposed algorithm which we use for comparison includes the face cluster labeling overhead as well. Finally, Figure 3.10 illustrates some dog and cat detection examples and Figure 3.12 presents an example scenario where the proposed cat face detection approach outperforms baseline as it limits the search ROI.

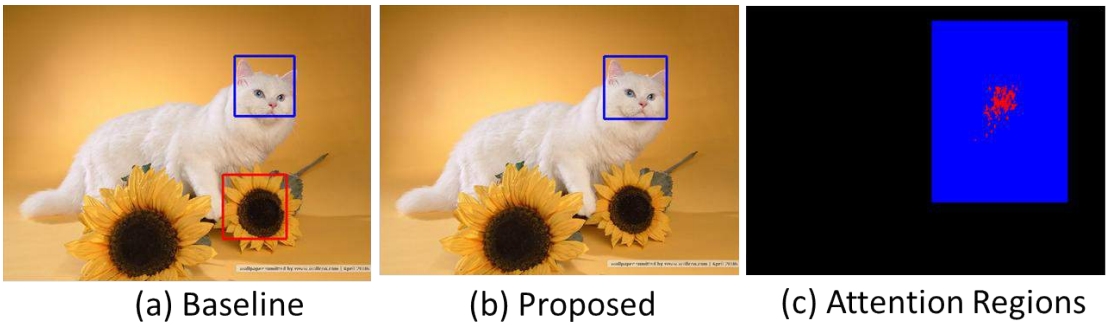



Figure 3.12: An example scenario where the head detector of the proposed approach (b) operating only in the attention region (c) marked in blue outperforms the baseline cat head detector (a). The baseline detector has a false detection as noticed in (a). Finally, red points in (c) denotes the cluster identified as face/head from which the blue attention region is constructed.

### 3.5.2 Detecting Text

Detecting text in natural scenes is an important problem for automatic navigation, robotics, mobile search and several other applications. Text detection in natural scenes is challenging as text is present in a wide variety of styles, fonts and shapes coupled with geometric distortions, varied lighting conditions and occlusions. Text detection approaches are divided into texture based and connected component (CC) based approaches. Texture based approaches typically learn the properties of text and background texture [23, 108] and classify image regions into text and non-text using sliding windows. Connected component (CC) based approaches [22, 82] group pixels which exhibit similar text properties. The grouping happens at multiple levels: character, word and sentence. This is followed by a geometric filtering technique which removes false positives. Stroke width transform (SWT) [32] is an elegant connected component

based approach which groups pixels based on the properties of the potential text stroke it belongs to. We utilize SWT as the baseline text detection algorithm as it obtained state-of-the-art results in the text detection datasets [61, 96] from which we obtained the images.

The first step of SWT is edge detection and the quality of edges primarily determine the final text detection performance [55]. The presence of several false edges especially in highly textured objects leads to false detections and therefore we propose an edge subset selection procedure from text priors obtained by labeling the eye tracking samples. A connected component edge map is obtained from the canny edges and we retain connected components that are sufficiently close to regions labeled as text. This is implemented by convolving the eye tracking samples using a Gaussian filter of variance 150 pixels (conservative selection) and obtaining a binary text attention map in the image plane by selecting regions which are above a threshold (0.02 in our case). In the following step, connected components of the edges which have an average text attention $> 0.4$ are retained for text detection.

The performance of the text detection is validated using standard precision-recall metrics popular in text detection literature[32]. Table 3.2 quantifies the improvements due to the proposed approach in precision and F-Measure of the text detector. *We notice significant gain in precision and F-Measure, about 37% and 15% respectively, compared to baseline SWT*. Table 3.2 also indicates that *we need to process only 34%*

| | Precision | Recall | F-Measure | Mean Edges |
|---|---|---|---|---|
| SWT | 0.436 | 0.669 | 0.530 | 6723 |
| Our Method | 0.599 | 0.655 | 0.625 | 19745 |

Table 3.2: Comparison of the performance of the proposed text detector with eye tracking prior and baseline SWT. There is significant gain in the precision (∼37% compared to baseline) for a small loss in recall (∼2%). This results in improved overall F-Measure.


Figure 3.13: Examples of images where the proposed text detection approach performs reliably.

*of the edges in the dataset which makes the proposed approach 2.82 times faster than baseline SWT*. We note that the time of the proposed algorithm which we use for comparison includes the text cluster labeling overhead as well. Figure 3.13 highlights some example detections from the proposed algorithm. Figure 3.14 compares some results of the proposed approach to baseline SWT and indicates the utility of the text attention map to limit the ROI for text detection. In summary, we obtain significantly better detector precision than baseline SWT in considerably lower detection time.

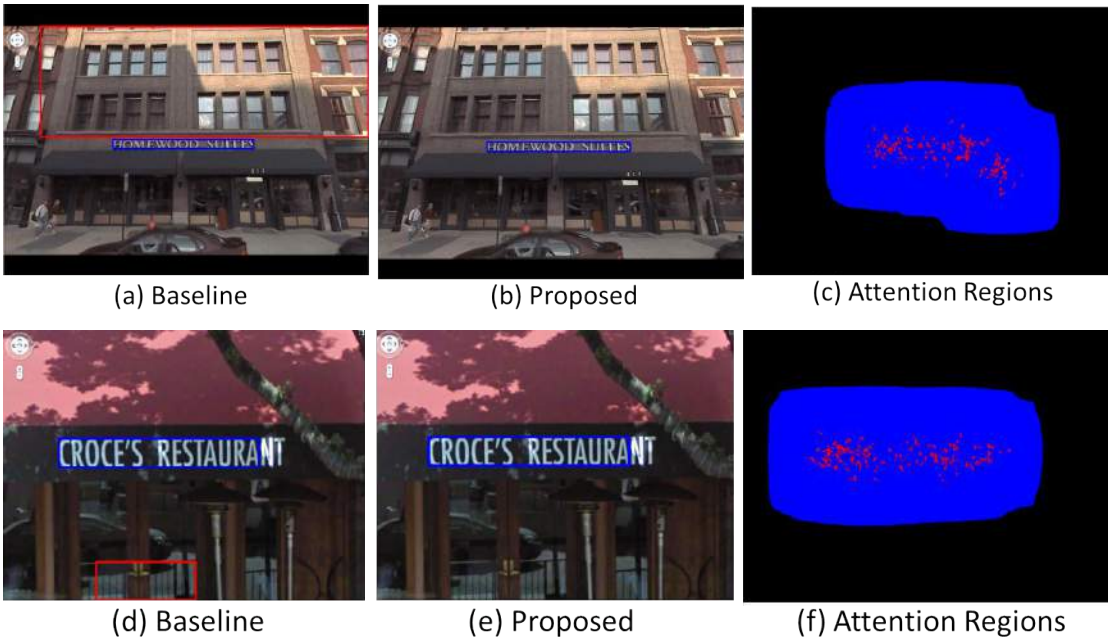



Figure 3.14: Two example scenarios ((a)-(c) and (d)-(f)) where SWT results ((a) and (d)) are outperformed by the proposed approach ((b) and (e)). The attention regions ((c) and (f)) shows the eye tracking samples classified as text in red and the ROI used by the text detector in blue. Therefore, as the false positive portion in SWT (red boxes in (a) and (d)) is removed by the generated text attention region, we obtain better detector precision in these images.

## 3.6 Discussion, Summary and Future Work

This work is the first attempt at interpreting image semantics from the manner in which multiple subjects look at these images in a free viewing task. Consequently, we generate semantic priors by analyzing eye tracking samples without image information. We focused on two semantic categories, faces and text, and collected a new eye tracking dataset. The dataset consists of 300 images with 15 subjects with specific focus on humans, dogs, cats and text in natural scenes. The eye tracking fixation samples

are clustered using mean-shift. Intra- and inter-cluster features are computed which eventually maps to a labeling problem using an MRF. The proposed approach obtains promising results in classifying face and text regions from background by only analyzing eye tracking samples. This information provides a very useful prior for challenging problems which require robust face and text detection. Finally the proposed semantic prior in conjunction with state-of-the-art detectors obtains faster detections and higher precision results for dog, cat and text detection problems compared to baseline.

The proposed approach also has a few limitations. If the face image almost occupies the entire screen, multiple clusters at different face parts will be formed and our dataset does not provide sufficient samples to model this behavior. Furthermore, if the image has a large number of text lines, the subjects do not have sufficient viewing time to gather all the information presented. This can be handled by allowing the subject to control viewing time. Both these issues will be addressed in future extensions of this work.

In addition, one can explore better localization of face and text regions for the detectors from the eye tracking information. Perhaps one could learn the relationship between the ground truth bounding boxes and the cluster properties. Additionally, an edge learning technique from the cluster labels for the text class could improve the proposed text detection algorithm. Finally, it would be interesting to investigate learning eye tracking priors for other semantic categories and over video sequences from

multiple subjects.

# Chapter 4

# Eye tracking assisted object extraction from videos

"What is important is not what you hear said, it's what you observe".

*Michael Connelly*

In the previous chapter we observed how eye tracking data from multiple subjects in a free viewing task can be used to localize specific objects in images. This idea can be extended to videos where eye tracking data is more efficient in annotating large number of frames in a relatively smaller duration. Therefore, in this work we propose an algorithm to extract objects from videos which attract visual attention. As human attention is naturally biased towards high level objects in visual scenes, this information

can be valuable to extract salient objects in a scene. The proposed algorithm extracts dominant visual tracks by a combination of mean-shift and Hungarian algorithm on eye tracking data from multiple subjects. These visual tracks guide a generic object search algorithm to get candidate object locations and extent in every frame. Further, we propose a novel multiple object extraction algorithm by constructing a spatio-temporal mixed graph over object candidates. The object extraction inference is obtained using binary linear integer programming. Finally, the object boundaries are refined using grabcut segmentation algorithm. The proposed technique outperforms state-of-the-art object segmentation using eye tracking prior and obtains favorable segmentation results over algorithms which do not utilize eye tracking data.

This work is organized as follows. We motivate the problem and discuss the related work in Section 4.1. In Section 4.2 we introduce the eye tracking dataset. Our algorithm to extract visual tracks from eye tracking data in described in Section 4.3. The multiple object extraction framework is also presented in this section following which in Section 4.4 we demonstrate the results of the proposed approach. Finally the conclusions and future work are discussed in Sections 4.5 and 4.6 respectively.

## 4.1 Introduction

Object extraction in videos is a challenging problem in computer vision. Automated extraction of objects in a video sequence can benefit several applications related to video annotation, compression, summarization, search and retrieval. A critical bottleneck in object extraction is defining the importance of objects in a video sequence. Several works in object extraction from videos have focussed on utilizing motion to determine the importance of objects in video sequence. These methods typically aim to extract a dominant object in the scene, where object importance is determined by motion. In [59] Lee et al. identify important motion segments representing an object and extrapolate the object of interest throughout the video frames. In [62] Ma et al. extracted objectness proposals from all video frames and identify the important object by connecting the proposals using mutual exclusiveness contraints. In recent work in [106] Zhang et al. proposed a framework to extract objects using objectness [31] and optical flow proposals and segment the key object by dynamic programming on a directed acyclic graph. They also indicate a technique to ensure robustness to broken object segments. All the aforementioned techniques utilize motion to define the importance of objects and can extract only a single object of interest from a video sequence. However, motion may not be a good metric to determine importance of objects in videos. For example in a video sequence where two subjects are having a conversation, the motion cues might be misleading. We note that extraction of salient objects in a scene can be

better understood by visual attention it attracts in a scene. Therefore, in this work we investigate the utility of eye tracking to extract multiple interesting objects in a scene.

Eye tracking data is biased towards high level semantics in static and dynamic scenes. Therefore, visual attention can provide a robust prior to assist multiple object extraction problem in video sequences. Recent advancements in eye tracking technology has opened up avenues to collect data without affecting the experience of the viewer. State-of-the-art eye trackers are affordable [4] and this has enabled large-scale collection of eye tracking data from multiple subjects. Multimedia content is typically viewed by a large number of people and collecting eye tracking data from a small fraction of the viewers can provide weak supervision to guide object extraction. Therefore, given a video sequence and eye tracking data from multiple subjects the objective is to extract relevant objects of interest which attract visual attention. A visual illustration of the proposed work is shown in Figure 4.1. Relevant to the proposed approach Mishra et al. [68] propose a segmentation using fixation approach which segments objects of interest given a single fixation point. They convert the image to polar coordinate space and graph cut segmentation in the polar coordinate space corresponds to object contour in the spatial domain. The approach was further extended using optical flow to segment a single object around a fixation point in a video sequence. The primary limitation of [68] is that they use a single fixation point and assume the fixation point is completely inside the object of interest. However, the assumption can be limiting as there is cali-

bration error in real eye tracking data especially when we have to extract small objects. Additionally, [76, 94] have proposed image segmentation algorithms using multiple fixations in order overcome some of the limitations of [68]. Recently, Papadopoulos et al. [73] explored the interesting problem of weakly annotating objects using eye tracking data to train object class detectors. The eye tracking annotations are used in the training phase to localize object bounding boxes which help train a deformable part model [35] based detector. The final detection performance is considerably lower than perfect ground truth annotations, however these annotations are obtained in about a sixth of the time required to hand annotate the bounding boxes which is encouraging. In [81] we propose an algorithm to extract face and text semantic priors using eye tracking data from multiple subjects and use this to enhance state-of-the-art object detectors. The algorithm is designed for images and is targeted for only face and text categories. Eye tracking based activity and action recognition techniques [92, 67] have also shown promise. In this work we propose an eye tracking assisted object extraction framework which is not restricted to specific object categories. The contributions of the proposed approach are as follows.

- A method to localize visual tracks from eye tracking data by solving a linear assignment problem, which coarsely corresponds to object locations in video sequences
- A novel object extraction framework guided by visual tracks, which extracts mul-

tiple objects in a spatio-temporal mixed graph by solving a binary integer linear program

- A novel eye tracking dataset on standard video sequences

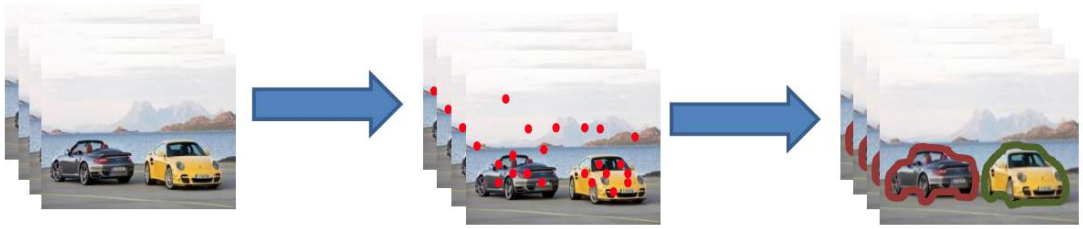


Figure 4.1: A simple illustration of the proposed problem. Given a video sequence, we collect eye tracking data in the sequence from multiple subjects and utilize this information to extract visually salient objects.

## 4.2 Eye tracking dataset on videos

We collected an eye tracking dataset on videos using Eyelink 1000 eye tracking device [3]. The users rest their head on a chin rest and the eye position is sampled by the eye tracker at 500 samples per second. Eye tracking data typically consists of fixations and saccades. The information gathering stage is encoded in the fixations and the saccades represent attention shift from one fixation to another. The eye tracker also segments the samples as fixations and saccades. The video dataset consists of 20 videos collected from SegTrack [91], GaTech [40] and Chen Xiph.org [21] datasets. The dataset consists of 1 to 4 dominant objects. The depicted scenes are obtained

from static and moving cameras with static and moving objects of interest. Figure 4.2 highlights examples video frames from input videos from the dataset. The videos were viewed by 21 subjects (between ages 21 and 35). The viewers sat 3 feet away from a 27 inch screen. The subjects were informed that it was a free viewing experiment and the data was collected without any apriori bias. This dataset can be downloaded from http://vision.ece.ucsb.edu/.

## 4.3 Proposed approach to extract objects from videos using eye tracking prior

In the following we will utilize eye tracking data as additional prior to improve object extraction from videos. An overview of the proposed approach is shown in Figure 4.3. The top row indicates the eye tracking processing steps to extract dominant visual tracks from eye tracking data from multiple subjects. The bottom row describes the visual track guided object extraction and novel multiple object extraction framework on a mixed graph. Finally the object boundaries are refined by segmentation using bounding box prior. The following sections provide a detailed description of the different modules which comprise the proposed framework.

Figure 4.2: Illustrates some example frames from the videos in the eye tracking dataset collected from Chen Xiph.org, GaTech and SegTrack datasets. We note that the dataset consists of single and multiple stationary and moving objects with moving and stationary backgrounds.

### 4.3.1 Eye tracking data processing to obtain dominant visual tracks

In order to extract dominant visual tracks from eye tracking data, we first introduce a simple pruning step to remove non-object eye tracking data. Eye tracking data is biased towards high level semantic objects and as described in Section 4.1 consists of fix-

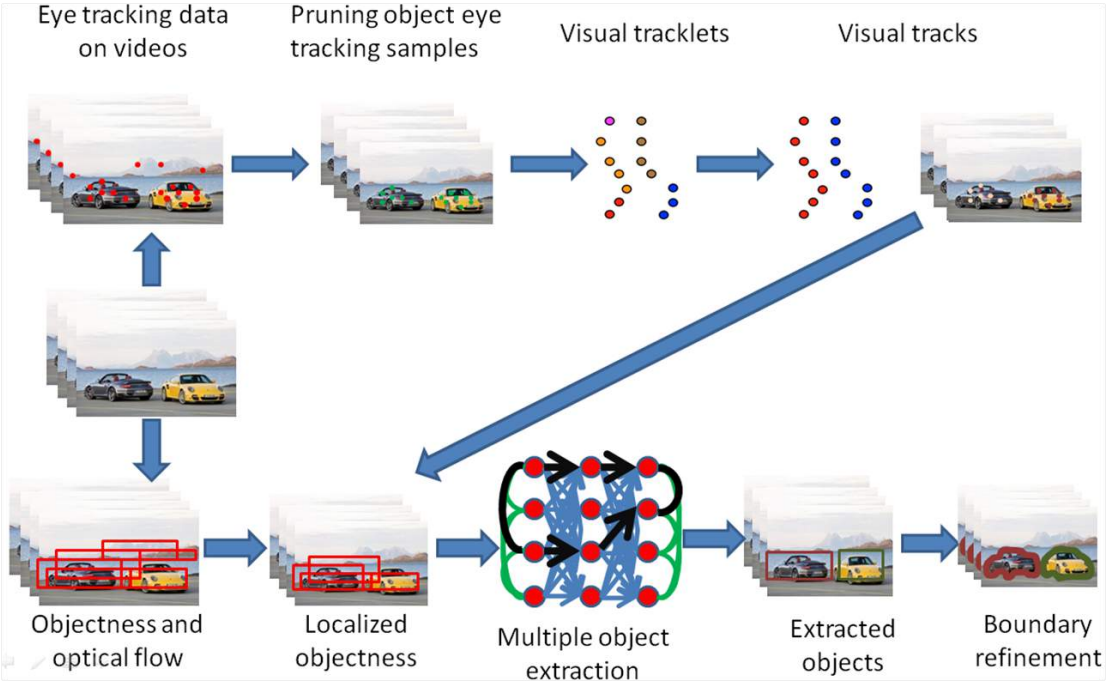



Figure 4.3: Block diagram of the proposed approach to extract multiple objects from videos using eye tracking prior. The top row indicates the eye tracking processing stage. The bottom row is the multiple object extraction framework guided by the visual tracks.

ations and saccades. The fixations represent the information gathering stage. Typically fixations are present in video regions where objects are present. However, saccades may or may not lie on objects in a video sequence. They lie on objects when the user is tracking a moving object and are called smooth pursuit (they are classified by the eye tracker as saccade) . However, saccades do not predominantly lie on any critical object when the user is shifting attention from one object to another. Therefore, we want to prune the saccades which indicate attention shift from one object to another. We utilize optical flow to determine the nature of the saccades and if the saccades lie in

the direction of optical flow we keep it for further processing otherwise we remove the saccade. The eye tracker extracts saccadic scanpaths $S_i$ which have individual samples $s_{ij}, j \in \{1...N\}$. Let $x_{ij}$ denote the pixel location of $s_{ij}$ in the image place. The optical flow at $x_{ij}$ be denoted by $O(x_{ij})$. The saccadic scanpath $S_i$ lies in the direction of optical flow if $\sum_j \frac{s_{ij}.O(x_{ij})}{|s_{ij}||O(x_{ij})|} < t$. This is illustrated in Figure 4.4.

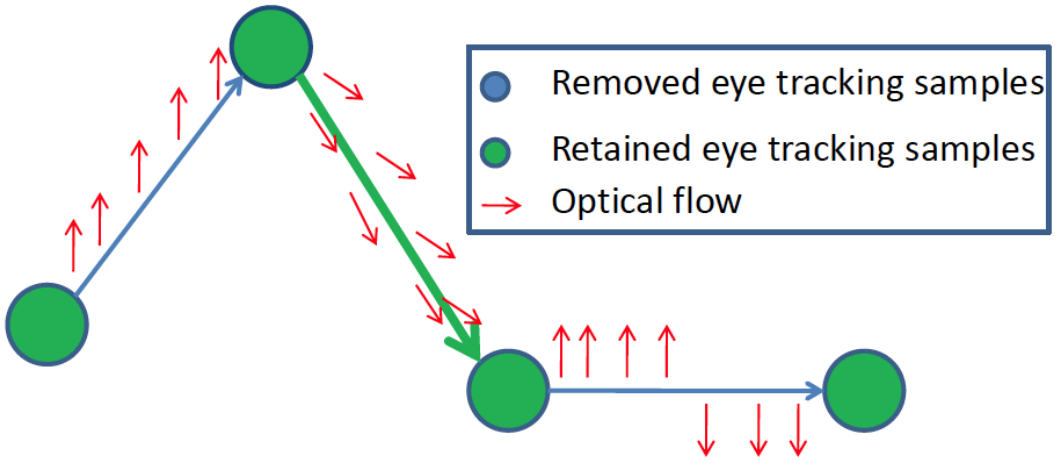



Figure 4.4: Illustrates that the saccades in the direction of optical flow are probable object saccades (smooth pursuit) and should be utilized along with the fixations for object localization. Saccades not in the direction of optical flow indicate attention shift from one object to another and can be pruned.

These pruned eye tracking samples are more probable to lie on objects in the videos compared to raw eye tracking samples. In the next stage, we associate these eye tracking samples to extract dominant visual tracks which coarsely corresponds to objects of interest in a video sequence. This is achieved by a two step hierarchical association process similar to [48]. First, the eye tracking samples are associated in a conservative

manner using mean shift clustering. This gives us tracklets representing eye tracking data over small segments of objects through the video sequence. In the next step the tracklets are associated to eventually represent dominant visual tracks.

Several association models have been proposed for multiple object tracking as reviewed in [54]. We use an approach similar to [48, 54] in which the authors jointly model the tracklet associations with the false alarm hypothesis. Let the individual tracklets of $\mathcal{T}$ be denoted by $\{T_1, T_2 ... T_N\}$. Similarly let the tracks of $\mathcal{S}$ be denoted by $\{S_1, S_2 ... S_M\}$.

Now, the association term is decomposed as

$$\begin{aligned} \mathcal{P}(\mathcal{S}|\mathcal{T}) &= \mathcal{P}(\mathcal{T}|\mathcal{S})\mathcal{P}(\mathcal{S}) \\ &= \prod_{T_k \in \mathcal{T}} \mathcal{P}(T_k|\mathcal{S}) \prod_{S_l \in \mathcal{S}} \mathcal{P}(S_l) \end{aligned} \tag{4.1}$$

Here we assume the likelihoods of the input tracklets are conditionally independent given $\mathcal{S}$ and the tracks $\{S_l\}$ are independent of each other.

A Bernoulli distribution is used to model the false alarm hypothesis of the tracklet using the detector precision denoted by $\beta$. Therefore, the likelihood of a tracklet is defined as

$$\mathcal{P}(T_k|\mathcal{S}) = \begin{cases} \mathcal{P}_+(T_k) = \beta^{|T_k|} & \text{if } \exists S_l \in \mathcal{S}, T_k \in S_l \\ \mathcal{P}_-(T_k) = (1-\beta)^{|T_k|} & \text{if } \forall S_l \in \mathcal{S}, T_k \notin S_l \end{cases} \tag{4.2}$$

where $|T_k|$ is the number of detections in $T_k$, and $\mathcal{P}_+(T_k)$ and $\mathcal{P}_-(T_k)$ are the likelihoods of $T_k$ being a true detection and a false alarm respectively.

The tracklet association priors in (4.1) are modeled as Markov Chains.

$$\mathcal{P}(S_l) = \mathcal{P}_{link}(T_{k_1}|T_{k_0})...\mathcal{P}_{link}(T_{k_{p_k}}|T_{k_{p_k-1}}) \tag{4.3}$$

where $p_k$ refers to the number of tracklets associated to form the track $S_k$. Basically, the association prior is a product of transition terms representing linkage probabilities between tracklets.

We note that $T_k$ cannot belong to more than one $S_l$. Thus (4.1) is rewritten as the following by inserting $\mathcal{P}_+(T_k)$ into its corresponding chain.

$$\mathcal{P}(\mathcal{S}|\mathcal{T}) = \prod_{\forall S_l \in \mathcal{S}, T_k \notin S_l} \mathcal{P}_-(T_k) \prod_{S_l \in \mathcal{S}} \Big[\mathcal{P}_+(T_{k_0}) \\ \mathcal{P}_{link}(T_{k_1}|T_{k_0})..\mathcal{P}_{link}(T_{k_{p_k}}|T_{k_{p_k-1}})\mathcal{P}_+(T_{k_{p_k}})\Big] \tag{4.4}$$

As we need to maximize (4.4), first we convert it into a cost function by taking negative logarithms. The cost described in (4.4) can be optimized by the Hungarian algorithm over tracklets similar to the one proposed in [48]. Here, a probabilistic cost is formulated to associate any two tracklets and to denote a tracklet as a false positive. In brief, to associate $n$ tracklets a $n \times n$ cost matrix is built with the non-diagonal entries denoting the tracklet association costs and the diagonals are the false positive costs. The optimal tracklet associations and false positive set which minimize the cost globally is obtained by the Hungarian assignment on this cost matrix. Therefore, the

joint cost matrix $\mathbf{C}_J$ of dimensions $n \times n$ to associate any two tracklets $T_p$ and $T_q$ is expressed as

$$\mathbf{C}_J(p,q) = \begin{cases} \ln \mathcal{P}_-(T_p) & \text{if } p = q \leq n \\ \ln \mathcal{P}_{link}(T_q|T_p) + 0.5[\ln \mathcal{P}_+(T_p^i) + \ln \mathcal{P}_+(T_q^i)] & \\ & \text{if } \ p,q \leq n \text{ and } p \neq q \\ -\infty & \text{otherwise} \end{cases} \tag{4.5}$$

The optimal tracks are obtained by the Hungarian algorithm on $\mathbf{C}_J$ which assigns every row to a unique column. If a tracklet is assigned to itself, it is a false positive tracklet and is removed from the dominant visual track list.

### 4.3.2 Visual track guided object extraction from videos

The visual tracks coarsely localizes interesting objects in a video sequence and thereby reduces the search space for important objects in the scene. Here, we note that the importance is determined by visual attention. Specifically visual tracks provide the following two critical pieces of information

- Number of visually salient objects in the scene
- Coarse spatial localization of the objects of interest

In this section we propose a novel principled framework to extract important objects of interest guided by the visual tracks. In order to determine object location in images,

we utilize the objectness measure to obtain objectness proposals. We notice that typical objectness provides several overlapping bounding boxes around an object of interest. Each bounding box is assigned an objectness score which indicates the score of the bounding box representing an object. We refine this score to reflect motion information by adding an additional term which measure optical flow magnitude contrast within and outside the bounding box. Let the optical flow magnitude sum within a bounding box $i$ and frame $f$ be $O^{if}_{in}$ and outside it be $O^{if}_{out}$. Then, the optical flow score is measured as $S^{if}_{opt} = 1 - e^{-\frac{(O^{if}_{in} - O^{if}_{out})^2}{\sigma_{opt}}}$. The overall combined objectness and optical flow score for bounding box $i$ in frame $f$ is a linear combination of individual scores and is given by $S^{if}_{comb} = S^{if}_{obj} + \alpha S^{if}_{opt}$.

Now given a set of bounding boxes in every frame, and the number of objects $k$ (number of visual tracks) we want to extract $k$ distinct objects from the video sequence. Each box has a unary score indicated by $S_{comb}$. In addition we also define pairwise costs across bounding box pairs in successive frames. This score is determined from overlap distance and color histogram distance between the two frames. Let $b^i_f$ and $b^j_{f+1}$ represent two bounding boxes in successive frames $f$ and $f+1$, then the pairwise score is represented as $S^{ijf}_{pair} = S^{ijf}_{overlap} + \beta S^{ijf}_{color}$, where $S^{ijf}_{overlap} = \frac{b^i_f \cap b^j_{f+1}}{b^i_f \cup b^j_{f+1}}$ and $S^{ijf}_{color} = 1 - e^{-\frac{(h^i_f - h^j_{f+1})^2}{\sigma_{color}}}$, where $h^i_f$ and $h^j_{f+1}$ are the color histograms of the bounding boxes in frames $f$ and $f+1$. Therefore, the overall combined unary and pairwise score is represented as $S^{ijf}_{overall} = S^{if}_{comb} + S^{i(f+1)}_{comb} + \gamma S^{ijf}_{pairwise}$. The overall temporal cost

between bounding boxes $b_f^i$ and $b_{f+1}^j$ is $C_{temp}^{ijf} = 1 - S_{overall}^{ijf}$.

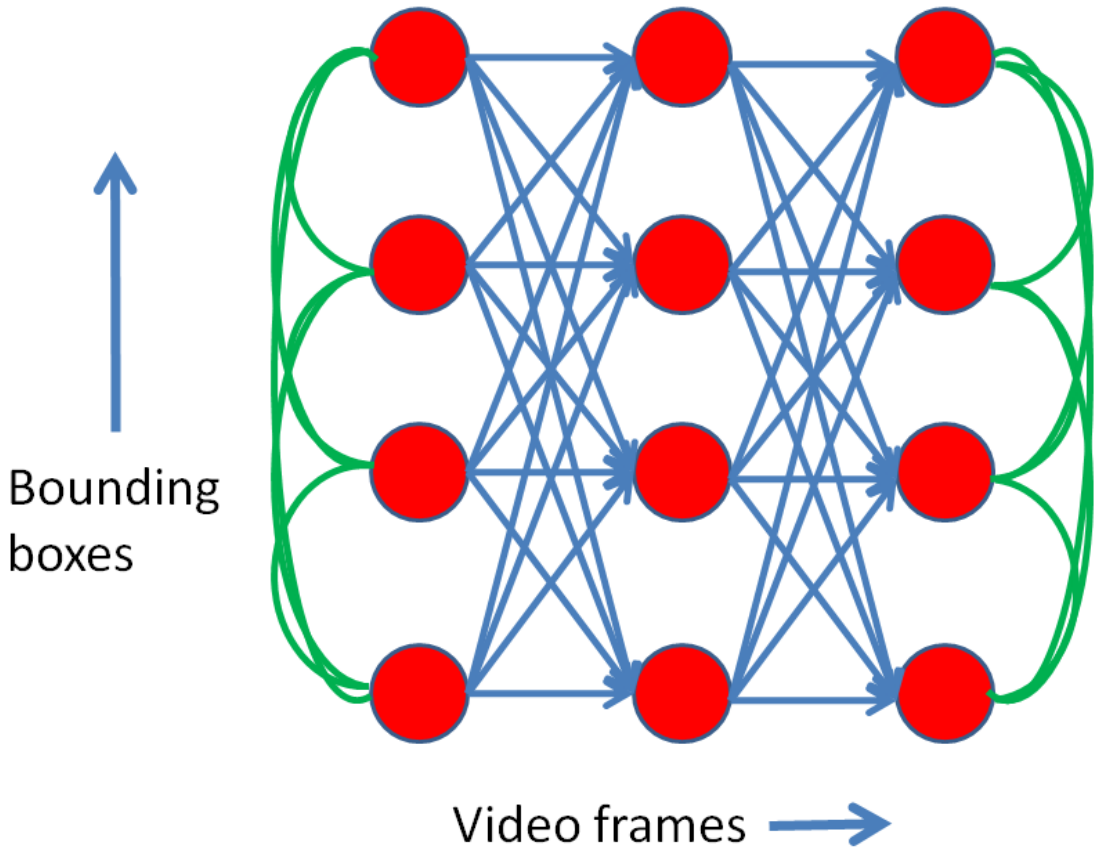



Figure 4.5: The spatio-temporal graph to extract multiple objects is highlighted here. The temporal costs shown in blue indicate inter-frame cost to connect a path through two bounding boxes in successive frames. The intra-frame spatial costs are indicated in green. They aim to penalize extraction of the same object in multiple paths.

We construct a graph using this cost and the aim is to extract $k$ paths through the graph which minimize the overall cost as shown in Figure 4.5. As objectness metric extracts multiple bounding boxes around an object of interest it is possible to extract the same object in several paths. In order to mitigate this we introduce spatial costs within

a frame. The spatial cost ensures that the same object is not extracted in multiple paths through the graph. The spatial cost associated with two bounding boxes $b_f^i$ and $b_f^j$ in frame $f$ is $C_{spatial}^{ijf} = \frac{b_f^i \cap b_f^j}{b_f^i \cup b_f^j}$. Therefore, the aim is to select paths which minimize the overall spatio-temporal cost. Let the decision variables for the temporal costs be denoted by $x_f^{ij}$ between bounding boxes $i$ in frame $f$ and $j$ in $f+1$. Also, let the decision variables for the spatial costs be denoted by $y_f^{ij}$ between bounding boxes $i$ and $j$ in frame $f$. Assuming the total number of frames is $F$, the optimization problem can be formulated as

$$\underset{x_f^{ij}, y_f^{ij}}{\text{minimize}} \sum_{i,j,f} C_{temp}^{ijf} x_f^{ij} + \sum_{i,j,f} C_{spatial}^{ijf} y_f^{ij} \text{ subject to}$$

$\sum_i x_f^{ij} = \sum_k x_f^{jk} \forall j, f$ : Conservation of flow constraint

$\sum_j x_0^{sj} = k$ : Flow from source node = $k$ to get $k$ distinct paths

$\sum_i x_F^{it} = k$ : Flow to terminal node = $k$, conservation of flow

$\sum_i x_f^{ij} = 1 \forall j, f$ : Two temporal paths to a node cannot be active

$y_f^{ij} = (\sum_k x_f^{ik})(\sum_k x_f^{jk}) \forall i, j$

Can be linearized as $y_f^{ij} \geq \sum_k x_f^{ik} + \sum_k x_f^{jk} - 1$

$x_f^{ij}$ and $y_f^{ij} \in \{0, 1\}$

This results in a binary integer linear program as the decision variables are binary and the constraints and the cost functions are linear. We utilize the GUROBI [42]

solver to get the optimal solution to the problem which eventually extracts $k$ distinct paths from the graph.

Finally, the bounding box based object regions are refined using grab cut segmentation. The initial bounding boxes extracted from the graph are iteratively refined in every frame individually using grabcut segmentation [79].

## 4.4 Experimental results

In this section we evaluate the performance of object extraction using the proposed approach. First we evaluate the performance of the visual track extraction module. For this purpose we annotated important objects in the scene and define an object to be important if it captures more than 20% of the visual attention from all the observers. The ground truth important objects are defined by the following algorithm. We assume to have an exhaustive set of ground truth object annotations in an image. We sort the annotations by size and select the annotation which has more than 20% of the attention in the sorted order. Once an object is identified, we remove the object and its attention from the pool and repeat the process using the subsequent objects. Using this technique in the proposed dataset we observed 31 objects in total. The proposed approach extracted 30 objects (visual tracks) with 2 false negatives and 1 false positive.

After extracting the visual tracks representing objects in the video sequence, we

want to segment the object of interest from the videos. Our graph based object extraction algorithm extracts bounding boxes representing important objects in the scene. This is processed by the grabcut based refinement technique to obtain a better contour accurate representation of the object. Our approach is compared with [68] to evaluate the capability of multiple object extraction using eye tracking prior. As [68] requires a unique fixation point per object, we extract the median of the visual track in every frame to provide the fixation point which provides the pivot for the segmentation. Our approach is evaluated using track level VOC score between the ground truth and extracted objects. Let the ground truth track be $b^{GT}$ and the proposed approach track be denoted by $b^{our}$, then the VOC score is calculated as $\frac{b^{GT} \cap b^{our}}{b^{GT} \cup b^{our}}$. A comparison of our multiple object extraction algorithm with [68] is highlighted in Table 4.1. We notice that the proposed approach outperforms [68], which is the state-of-the-art in eye tracking assisted object extraction algorithm by a significant margin.

| | Our algorithm without Eye Tracking Data | Visual Tracks Only | Active Segmentation [68] | Our algorithm bounding boxes | Our algorithm with grabcut |
|---|---|---|---|---|---|
| Average VOC score | 0.21 | 0.23 | 0.38 | 0.37 | 0.46 |

Table 4.1: Comparison of the performance of our multiple object extraction algorithm with active segmentation using fixations [68]. We also selectively compare the performance of different sub-blocks of our model. We notice that both the object extraction module and eye tracking data contribute equally to extract objects which attract visual attention.

We also want to understand the role of eye tracking and object extraction module

individually to localize objects in a video. For this purpose we selected bounding boxes around every visual track by using $\mu \pm 2\sigma$, where $\mu$ and $\sigma$ are the mean and variance of the visual track in every frame. These bounding boxes represent eye tracking based bounding boxes ignoring visual information from the video sequence. Additionally, the multiple object extraction module is individually run on the video sequence without the eye tracking based localization prior to quantify the performance of the multiple object extraction framework without utilizing the eye tracking data. However, we utilize the number of visual tracks to extract $k$ objects from the video sequence. We notice in Table 4.1 that the proposed approach outperforms individual eye tracking and object extraction methods. Some example object extraction results using the proposed approach is shown in Figure 4.6. After applying grabcut, the final multiple object video segmentation results on a few example videos is shown in Figure 4.7. Finally, we also illustrate some example results from [68] in Figure 4.8. We notice that their algorithm is highly sensitive to the location of the fixation and noise in fixation localization can severely affect their performance.

Figure 4.6: Shows example results using the proposed approach to extract multiple objects represented by bounding boxes. We see the proposed approach is able to localize different visually salient objects in the video sequences with reasonable accuracy.

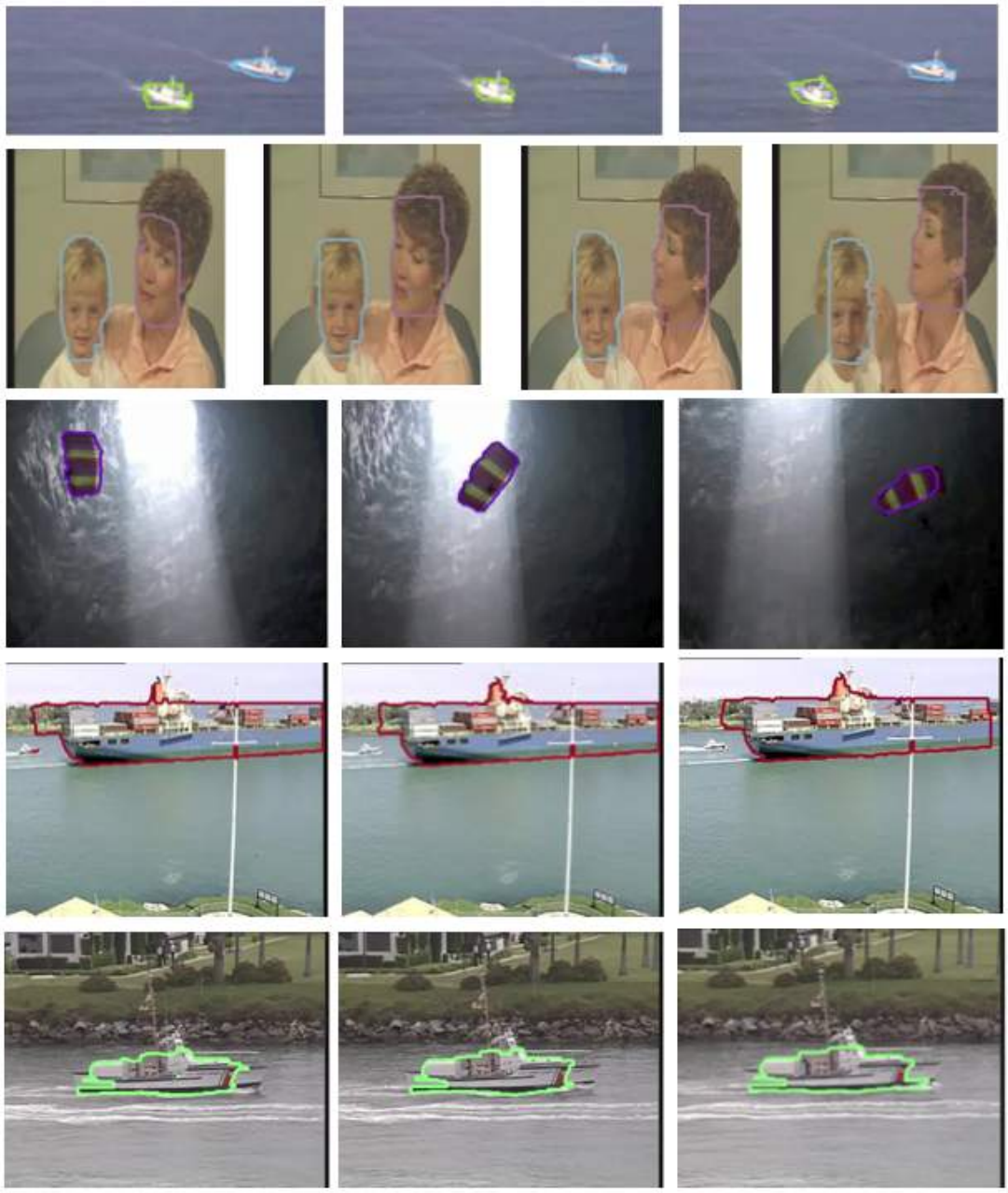


Figure 4.7: Shows example results from the proposed approach after applying grabcut based video segmentation to the extracted multiple object bounding boxes. We see the proposed approach is able to segment multiple objects in the video sequences with reasonable accuracy.

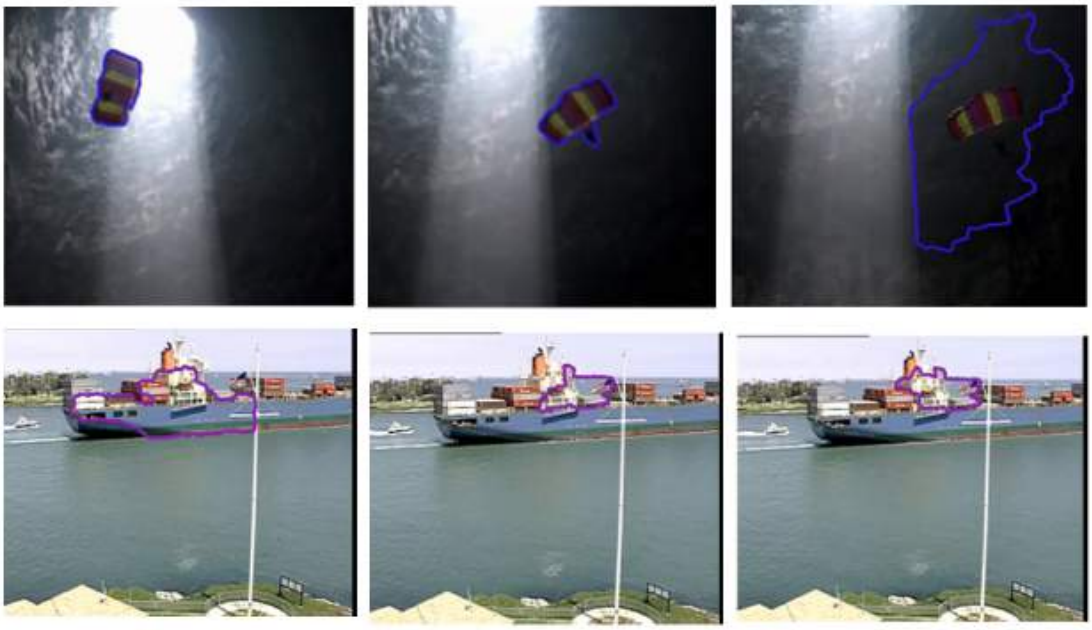


Figure 4.8: Shows some segmentation results using [68]. We notice that in the top row, when the fixation slightly positioned outside the object of interest, [68] breaks down. In addition, the algorithm suffers from similar issues in the bottom row as well as not being robust to the presence of the occluding pole.

## 4.5 Summary

Human visual attention is significantly biased towards high level semantic objects in visual scenes. Therefore, this information can be extremely useful to extract important objects in videos. Recent advances in eye tracking technology has enabled collection of eye tracking data from several subjects on a large scale. Multimedia applications can significantly benefit from the availability of such technology. This work proposes a novel framework for multiple object extraction using eye tracking data. The algorithm first clusters the eye tracking data using 3D mean shift to obtain visual tracklets, which are in turn associated to get visual tracks which coarsely localizes objects in a video sequence. The number of visual tracks indicates the number of visually significant objects and the extent of the track can be utilized to reduce the search space for objects. The visual track guided object search provides object proposals in every frame using objectness measure. Further, this information is used to build a spatio-temporal mixed graph and we extract paths representing objects from this graph by inference using binary integer linear programming. The extracted bounding box based objects are refined using grabcut segmentation to get object contour based segmentation.

The proposed approach outperforms the state-of-the-art object extraction using eye tracking fixation information. In addition we note that proposed combined framework which utilized both the object extraction and eye tracking information outperforms the individual modes of object localization.

## 4.6 Future Work

The proposed work is the first attempt to tackle object extraction from videos guided by eye tracking data. Several other problems in computer vision can benefit from the presence of eye tracking data. It would be interesting to explore the importance of eye tracking in image retrieval and activity recognition problems. Also, single subject eye tracking guided algorithms need further research as they will enable applications beyond multimedia where it can be combined with wearable technology. With respect to the proposed approach, in the future it would be interesting to explore the importance of how the number of eye tracking subjects affects object extraction performance.

# Chapter 5

# Saliency enhanced computer vision

"The eye sees only what the mind is prepared to comprehend."

*Roberston Davies*

The previous chapters utilized eye tracking to improve object detection in images and videos. In this chapter, we utilize saliency maps which mimic human attention to guide search for text. Humans have a remarkable ability to quickly discern regions containing text from other noisy regions in images. The primary contribution is to learn a model to mimic this behavior and aid text detection algorithms. The proposed approach utilizes multiple low level visual features which signify visually salient regions and learns a model to eventually provide a text attention map which indicates potential text regions in images. In the next stage, a text detector using stroke width

transform only focuses on these selective image regions achieving dual benefits of reduced computation time and better detection performance. Experimental results on the ICDAR 2003 text detection dataset demonstrate that the proposed method outperforms the baseline implementation of stroke width transform, and the generated text attention maps compare favorably with human fixation maps on text images. This work is organized as follows. In Section 5.1 we introduce the problem and review the related literature. A brief background of the text detection algorithm adopted in this work is presented in Section 5.2. In Section 5.3 we present our algorithm to learn text attention maps from visual saliency algorithms.In the following Section 5.4 we demonstrate the results of the proposed approach. Finally the conclusions and future work are discussed in Section 5.5.

## 5.1 Introduction

Detecting text in natural scenes is an important problem for automatic navigation, robotics, mobile search and several other applications. Text detection in natural scenes is challenging as text is present in a wide variety of styles, fonts and shapes coupled with geometric distortions, varied lighting conditions and occlusions. Text detection techniques can be broadly classified into two categories: texture based approaches and connected component based approaches. Texture based approaches learn the texture

differences between background and text regions. Image filtering techniques like Discrete Cosine transform [108] and Wavelet transforms [103] and Gabor filters [49] are commonly employed to represent the texture of text. These approaches typically use sliding windows and classify local image regions as text or non-text.

The second class of connected component (CC) based approaches are motivated by grouping pixels which exhibit similar text properties. The grouping happens at multiple levels : character, word and sentence. This is followed by a geometric filtering technique which removes false positives. Shivkumara et al. [82] proposed a CC approach in the Fourier-Laplace domain and geometric filtering using text straightness and edge density. Chen et al. [22] illustrated a CC based approach using Maximally Stable Extremal Regions (MSER). The popular Stroke Width Transform (SWT) [32] formulated by Ephstein et al. is also a CC based approach.


Figure 5.1: Left to right: 1. Input image. 2. Text attention map derived using visual attention features. 3. The text detection output indicated by the blue rectangle. Best viewed in color.

SWT is an elegant approach to detect text. However, its performance heavily relies

on the quality of edges which drive the transform computation. We propose a visual attention inspired solution to prune the search space of the SWT detector closer to text edges. In a free viewing task, human visual attention is heavily biased towards text regions [17] which have specific low level attention properties. Therefore, bottom up visual attention models which are designed to mimic human attention provide a useful prior for text detection. Given a set of training images, we compute several low level visual saliency maps, and train a classifier to understand both correctly and incorrectly labelled text and non-text regions provided by SWT detector. In a new test image, we use this classifier to produce a text attention map and SWT based text detection search is restricted to regions highlighted by this map improving both the speed and robustness of the detector. An example text detection obtained using our approach is shown in Figure 5.1.

In [88] Sun et al. also proposed a visual attention based text extraction approach based on Itti and Koch maps [51]. They used a predefined linear combination of the intensity, color and orientation channels to derive a map which filters false text blocks from potential character areas obtained by simple connected component analysis (CCA). This approach has several drawbacks. First, CCA based text detection is unreliable in the presence of noisy edges. Further, the weights for different features cannot be precomputed as in [88] when the number of bottom up features is large and finally [88] does not provide text attention map which prunes the detector search space.

Our approach overcomes the limitations of [88] by enhancing the state of the art SWT detector. The primary contributions of our work are

- Learning a model to derive text attention maps for images from multiple bottom-up saliency features. These maps compare favorably to human fixations in text images.
- Utilizing the learnt text attention map to improve the speed and accuracy of the stroke width transform algorithm.

## 5.2 Background: Stroke Width Transform

The proposed work aims to improve Stroke Width Transform (SWT) algorithm. SWT is a CC based approach with four stages, stroke width computation, character level grouping, geometric filtering and text line grouping. These stages are briefly described below.

Stroke Width Computation: Given an image, a corresponding edge map is computed using Canny edge detector. In addition, a gradient map is also obtained. From every edge pixel, rays are projected in the direction of the gradient until it encounters another edge pixel with an opposing gradient which is in the interval $[+\frac{\pi}{6}, -\frac{\pi}{6}]$ from the original gradient direction. If this condition is satisfied, pixels traced in this process potentially belong to the cross section of a stroke and are labelled as stroke pixels with width value

equal to the euclidean distance between the two edges. If an opposing gradient is not encountered, the ray is discarded or no stroke value is assigned to the pixels traced in that process. However, this approach fails in the intersection of multiple strokes like the junction present in ”T” as opposing gradient is absent in edges belonging to the junction. To fix this problem, a second iteration is performed along the edge pixels. Here, the discarded pixels are marked as strokes if more than a significant portion of these pixels have non-zero stroke width value from the first iteration. Finally, we obtain a map with potential strokes. To detect both bright and dark strokes, this algorithm is executed twice, in both the positive and negative gradient direction.

Character level grouping: In this stage similar strokes widths are grouped into characters using a modified connected component algorithm. This algorithm ensures grouping of two neighboring pixels if their stroke width ratio is in the range $[3, \frac{1}{3}]$.

Geometric filtering: Detected character regions which do not satisfy certain geometric properties related to aspect ratio, median stroke width and size of the connected components are discarded.

Text line grouping: Characters which have similar stroke widths, letter widths, height and spaces between letters and words are grouped to obtain text lines. A text line must have minimum three characters to suppress false detections.

For a detailed version of this algorithm we refer to [32]. A visual example describing the steps in SWT is shown in Figure 5.2 code is not publicly available and we

implemented our version of the algorithm.


Figure 5.2: Left to right: 1. The input image. 2. Stroke Width Transform Image. 3. Connected components and geometric filtering. 4. Final Detections (blue boxes). Best viewed in color.

## 5.3 The Proposed Approach to learn text attention maps

The performance of SWT significantly depends on the quality of edges extracted from images. Typically, highly textured edges from trees, brick walls and other natural structures reduce the precision of SWT detector as it is prone to false positive detections in those regions. To overcome this problem, we develop an edge subset selection procedure which reliably detects *text edges*. Given a set of edges $\mathcal{E}$ in an image, we want to select a subset of edges $\mathcal{E}'$ which improves the SWT detector. Mathematically we want to obtain

$$\arg\max_{\mathcal{E}'} q_{SWT}^{\mathcal{E}'} \quad \text{S.T} \quad \mathcal{E}' \subseteq \mathcal{E} \tag{5.1}$$

where $q_{SWT}^{\mathcal{E}}$ is a quality measure of SWT detector using edges $\mathcal{E}$.

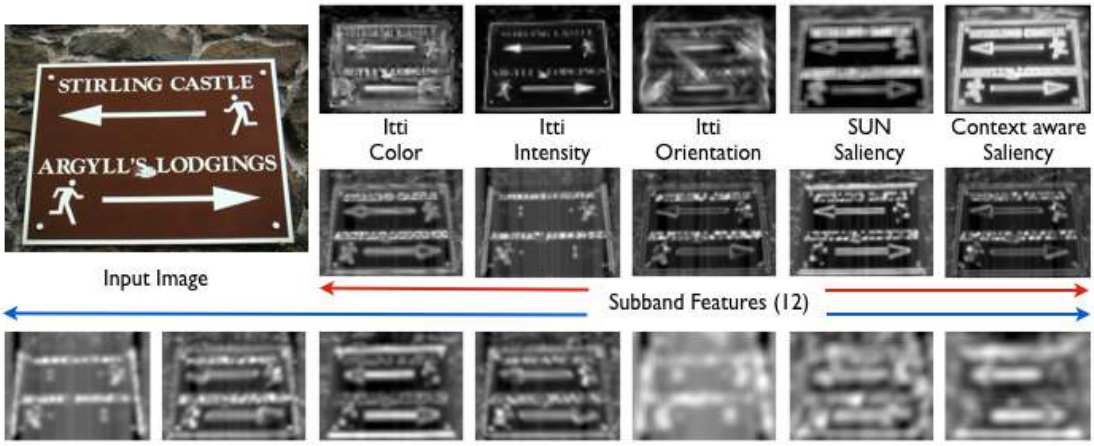



Figure 5.3: Example of visual attention features computed in an image.

### 5.3.1 Learning

The best $\mathcal{E}'$ would correspond to a subset of edges which belong to text. As humans are adept at text detection, biologically motivated low level visual attention features which mimic human attention provide a useful prior for text boundary detection. Therefore, to approximate (5.1), we propose a learning based algorithm which estimates a mapping from these multiple low level saliency maps to text regions in an image for removing distracting edges.

### 5.3.2 Features

The following low level features are used in our algorithm:

Itti and Koch Saliency map: This early saliency model [51] is motivated by linear filter-

ing and center surround operations and biologically motivated normalization provides intensity, color and contrast channels which we use in our model. This approach was primarily motivated for rapid analysis of visual scenes.

Context Aware Saliency Map: This approach [39] builds a mathematical model to the principles of human visual attention supported by psychological evidence which includes local global scale saliency, multi-scale saliency enhancement, immediate context inclusion, center prior and high level factors. This approach extracts salient objects together with parts of the discourse that surrounds them that can shed light on the meaning of the image.

Steerable pyramid features: The local energy of steerable pyramid filters [84] are correlated to visual attention. We use the features extracted from the pyramid subbands in four orientations and three scales similar to [53]. This combination provides 12 attention maps.

SUN Saliency map: Saliency Using Natural statistics [107] provides a map utilizing top-down and bottom-up information. This approach uses self information of visual features and pointwise mutual information between features and target during target search process.

We examined the utility of other saliency maps [53, 71, 46, 43, 56] which are effective in predicting human eye movements in natural images, however their text specificity was not suitable for our model, primarily attributed to the center bias prominent in

these saliency maps. In total we have 17 attention maps and an example of the different extracted features are shown in Figure 5.3.

Given an image, we want to learn a binary map which highlights image regions which have high probability of text using features signifying visual attention. This map is called a text attention map, obtained by training a classifier to understand a mapping from attention features to text regions in images. Given a training set, we use SWT to extract character regions in all the images. Using the ground truth labels, we obtain sufficient true and false positive character regions.

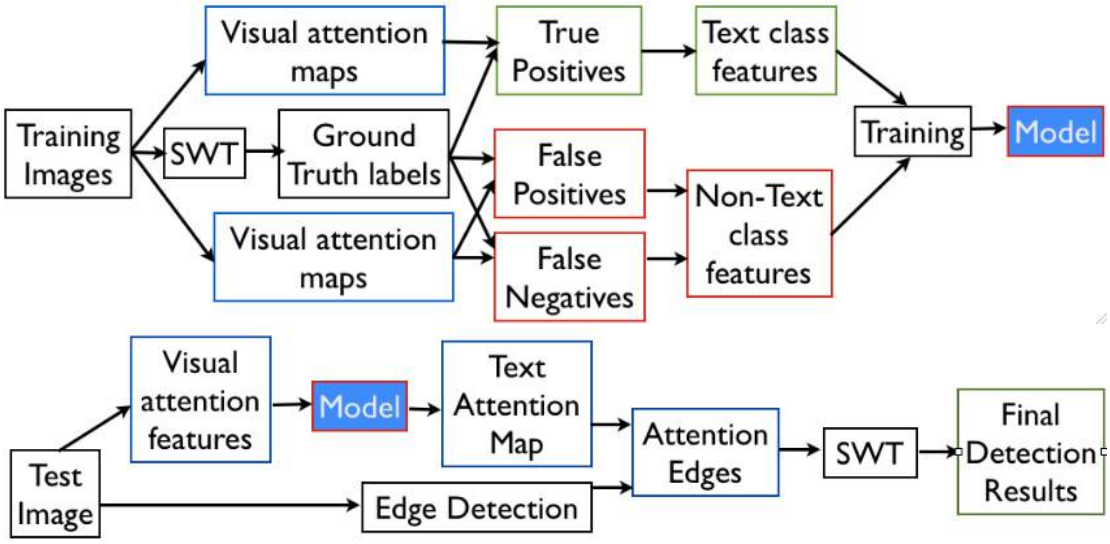



Figure 5.4: Block diagram of the training(top) and test (bottom)modules of the visual attention based learning paradigm.

A subset of pixels from true positive character regions are selected for training the text class. We also note that non-text class consists of equal number of pixels from false

positives and true negatives. This procedure ensures the training set for non-text class consists of sufficient examples where SWT usually provides false positives enabling the text attention map to correct SWT mistakes. Next, we learn a model to predict these text and non-text regions using visual attention maps at these selected pixel locations. Given a new image, this model (classifier) generates a corresponding text attention map by classifying every pixel as text or non-text and SWT based text detector only concentrates on regions classified as text. It offers dual benefits of lower computation time and higher precision. Further, the edges contained in these text attention maps approximate the edge subset selection problem (5.1). A block diagram of our framework is illustrated in Figure 5.4.

## 5.4 Experiments and Results

We perform two separate experiments to validate the effectiveness of the derived text attention maps. In the first part we compare the text attention maps to eye fixation data in the MIT eye tracking dataset [53]. In the second experiment, we aim to improve the detection performance of stroke width transform algorithm.

### 5.4.1 Dataset and Setup

The ICDAR 2003 [61] text detection dataset is used to evaluate our algorithm. The dataset consists of 258 training images and 251 test images with challenging text present in various fonts, sizes, backgrounds, transparency, non-planar surfaces and reflections with word level ground truth annotations. As SWT is originally designed to capture text line groups, words in a sentence are combined to obtain text line level annotations for training and testing. During training, we utilize SWT and obtain true and false positive character regions. Next, all the training maps are resized such that the largest dimension consists of 200 pixels while maintaining the original aspect ratio. From every resized map, we randomly sample $12\%$ of true positive locations for the text class and $7.4\%$ of false positives and $0.3\%$ of true negatives for non-text class. This gives about 40000 training samples per class (equal false positives and true negatives for non-text class). After training a model according to Section 5.3, for every test image (after resizing it in the same manner) we classify each pixel and obtain a text attention map by thresholding every pixel whose posterior probability of belonging to text class$>$ 0.35. This conservative threshold ensures most of the text regions are preserved in the map allowing some false non-text regions too. In the following stage, the SWT algorithm operates only on these regions for text detection. In practice we obtain a connected component canny edge map and every connected component which has more than $80\%$ attention edges is selected for SWT based text detection.

Algorithms 2 and 3 provide a step-by-step rundown of our training and test setup. We briefly explain the steps involved in the algorithms. In the training phase, for each training image, we compute the SWT ($\mathcal{SW}_i$) detections and obtains the true positives ($\mathcal{TP}_i^{sub}$) , false positives ($\mathcal{FP}_i^{sub}$) and true negative ($\mathcal{TN}_i^{sub}$) detections. This enables learning a positive training set (train$^+$) from the true positives and negative training set (train$^-$) from false positives and true negatives. A classifier ($\mathcal{C}$) is trained using this set. In the test phase, for each test image $\mathcal{I}$, the text attention value is computed per pixel using the learnt classifier $\mathcal{C}$. The attention map ($\mathcal{A}$) guides the selection of the edges for SWT.

### 5.4.2 Comparison to Human Fixations

The proposed approach to obtain text attention maps was motivated to mimic the manner in which humans viewed text images. To test that theory, we collected a set of text images from MIT eye tracking dataset [53] and compared the text attention map generated by our algorithm to the Gaussian smoothed human fixation map. Figure 5.5 illustrates a few example images with their corresponding human and text attention map. We notice that our text attention maps significantly correlate well with human attention maps for the specific class of text images.


Figure 5.5: Left column shows the input image, center column corresponds to human fixation map and the right column illustrates the proposed text attention map. The text attention maps are similar to human fixation map on text centric images. Note that eye fixations only includes foveal or central vision and peripheral vision is not captured. Therefore, as row 1 and 3 only have a single word, eye tracking results are biased towards the center of the word and therefore does not entirely overlap with our text attention map. Moreover, the text attention maps reliably localize the text regions.

**Data**: Input Images $\{\mathcal{I}_i\}$ and binary ground truth labels $\{\mathcal{L}_i\}$, $i \in [1, N]$

**Result**: Classifier Model $\mathcal{C}$

initialization $tp$=0.12, $fp$=0.074, $fn$=0.003;

**for** *i=1*$\rightarrow N$ **do**

$\mathcal{F}_i$ = features($\mathcal{I}_i$);

$\mathcal{SW}_i$ = Stroke Width Image($\mathcal{I}_i$);

$\mathcal{T}_i$ = Binary Mask($\mathcal{SW}_i$): Binary mask of character regions;

$\mathcal{TP}_i$=$\mathcal{F}_i(\mathcal{T}_i \bigodot \mathcal{L}_i)$: True Positives;

$\mathcal{FP}_i$=$\mathcal{F}_i(\mathcal{T}_i \bigodot (1 - \mathcal{L}_i))$: False Positives;

$\mathcal{TN}_i$=$\mathcal{F}_i((1 - \mathcal{T}_i) \bigodot (1 - \mathcal{L}_i))$: True Negatives;

$\mathcal{TP}_i^{sub}$ = Rand. Subset($\mathcal{TP}_i$) S.T $|\mathcal{TP}_i^{sub}| = \lfloor (|\mathcal{TP}_i|tp) \rfloor$;

$\mathcal{FP}_i^{sub}$ = Rand. Subset($\mathcal{FP}_i$) S.T $|\mathcal{FP}_i^{sub}| = \lfloor (|\mathcal{FP}_i|fp) \rfloor$;

$\mathcal{TN}_i^{sub}$ = Rand. Subset($\mathcal{TN}_i$) S.T $|\mathcal{TN}_i^{sub}| = \lfloor (|\mathcal{TN}_i|fn) \rfloor$;

**end**

train$^+$ = $\bigcup_i \mathcal{TP}_i^{sub}$;

train$^-$ = $(\bigcup_i \mathcal{FP}_i^{sub}) \cup (\bigcup_i \mathcal{TN}_i^{sub})$;

$\mathcal{C}$ = Classifier(train$^+$,train$^-$);

**Algorithm 2:** Training algorithm

**Data**: Test Image $\mathcal{I}$, Edgemap $\mathcal{E}$, Classifier $\mathcal{C}$, Connected Component Edges $\mathbf{C}^E$

**Result**: Text Attention Map $\mathcal{A}$, Attention Edges $\mathcal{E}'$ Detections $\mathcal{D}$

initialization $\mathcal{E}' = \emptyset$;

$\mathcal{F}$ = features($\mathcal{I}$); posterior = $\mathcal{C}(\mathcal{F})$;

**for** *i,j* $\in$ *[row,col]* **do**

$$\mathcal{A}(i,j) = \begin{cases} 1 & \text{posterior} > 0.35 \\ 0 & \text{else} \end{cases}$$

**end**

**for** *each c* $\in \mathbf{C}^E$ **do**

**if** $\frac{\sum_{p \in P} c(p)\mathcal{A}(p)}{|c|} > 0.8$ **then**

$\mathcal{E}' = \mathcal{E}' \cup c$

**end**

**end**

$\mathcal{D}$ = SWT($\mathcal{E}'$) : Stroke Width Transform on $\mathcal{E}'$

**Algorithm 3:** Testing algorithm

### 5.4.3 Text Detection Results

The output of text detection algorithm are a set of rectangles denoting text lines. These rectangles are matched to the ground truth rectangles representing text lines. A match score $m$, between two rectangles is determined as the intersection area divided by the union area. This quantity is 1 for identical rectangles and 0 for non-overlapping ones. For a given rectangle $t$ the best matching rectangle $m_b$ in a set of rectangles $\mathcal{T}$ is defined by $m_b(t, \mathcal{T}) = \max\{m(t, t')|t' \in \mathcal{T}\}$. This leads us to the definitions of Precision and Recall as Precision $= \frac{\sum_{t_e \in \mathcal{E}} m_b(t_e, \mathcal{G})}{|\mathcal{E}|}$ and Recall $= \frac{\sum_{t_g \in \mathcal{G}} m_b(t_g, \mathcal{E})}{|\mathcal{G}|}$. Here, $\mathcal{G}$ and $\mathcal{E}$ are the sets of ground truth and estimated rectangles respectively. The precision and recall are combined to a single quantity called $f$ measure which is defined as $f = \frac{1}{\frac{\alpha}{\text{Precision}} + \frac{1-\alpha}{\text{Recall}}}$. Typically $\alpha$ is set to 0.5.

| | Precision | Recall | $f$ Measure | Median Edges |
|---|---|---|---|---|
| SWT | 0.613 | 0.721 | 0.664 | 12723 |
| Our Method | 0.720 | 0.727 | 0.724 | 19745 |

Table 5.1: Comparison of the performance of our algorithm and SWT

First, in the training phase we used three classifiers: SVM with Radial Basis Function(RBF) Kernel [19], Lib-linear SVM [33] and Linear Discriminant Analysis based classifier (LDA) [8]. SVM with RBF kernel was able to learn a better model to predict text regions on a validation set, than the linear classifiers as it obtained 86.3% accuracy


Figure 5.6: Example detections (blue boxes) in images from ICDAR dataset. Best viewed in color.

compared to 78.3% and 77.1% by Lib-linear SVM and LDA respectively. This validation is a significant step as it provides evidence that bottom up visual attention based features can be used to understand text regions in images. Further, in the test stage, SVM with RBF kernel is used to compare our approach to SWT.

In the test phase, the proposed approach using SVM+RBF kernel obtains significantly better precision than baseline SWT and therefore $f$ measure of our algorithm outperforms baseline SWT by 9.04% as indicated in Table 5.1. The text attention map is also able to remove a significant portion of false positive edges (about 55% from


Figure 5.7: Illustrates two example scenarios where our algorithm (left) outperforms SWT (center). The text attention maps (right) clearly ignores regions where SWT detects false positives. The detections are shown in blue rectangles. Best viewed in color.

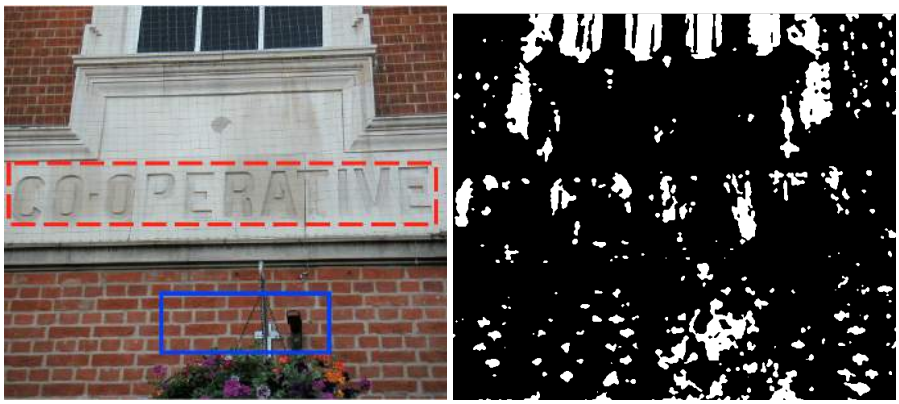



Figure 5.8: An example image (left) where the proposed algorithm fails and the corresponding attention map (right). In this image the background is very similar to text region, hence, the text attention map fails to localize the text region. The missed detections are shown in red rectangles and the false positive detections in blue rectangles. Best viewed in color.

Table 1) and our text attention coupled with SWT is 30% faster than baseline SWT. Figure 5.6 shows some example detections obtained from our algorithm. We are able to reject textured regions such as trees and bricks and is able to reliably detect text even in the presence of reflection and background clutter. Figure 5.7 highlights some visual examples where our method provides better detection results than SWT. The derived attention maps for these images indicate that edges corresponding to low contrast background regions (especially bricks) are ignored by the text attention maps leading to improved detection accuracy. Finally, Figure 5.8 shows an example where our approach fails to detect the text region in the image. The attention map in Figure 5.8 ignores the text region as it blends in with the surrounding background which caused the missed detection.

## 5.5 Summary

We have proposed a novel learning based framework to obtain text attention maps for images. These text attention maps prune the search space for SWT based detection. The overall pipeline significantly improves the precision of the SWT detector and also reduces the computation time. However, in regions where the text blends with the background our approach fails to detect the text. In addition, our attention maps resemble human attention maps in text images without multiple distractor elements. In the future

one can explore the possibility of adapting a learnt visual attention model to provide text attention maps instead of learning it ab initio from saliency maps.

# Chapter 6

# Role of scene and camera context in visual attention modeling

> ”Attention is the rarest and purest form of generosity.”
>
> *Simone Weil*

The previous chapters utilized eye tracking and saliency which mimic human attention to improve object detection algorithms. In this chapter, we tackle the inverse problem of developing better algorithms to predict where people look using advancements in object detection and other computer vision techniques. We first analyse the manner in which visual attention changes when camera focus changes. In order to understand this we utilize an image gradient based approach to detect in-focus and out-

of-focus regions in images. The analysis lays emphasis on the consistency of fixations across images when camera focus changes. We also propose a visual attention model to predict regions which humans typically fixate on. The model utilizes several low, mid, high and scene context features and utilizes a regression algorithm to predict the saliency in a test image. This model outperforms other state-of-the-art saliency and visual attention models in this light field eye tracking dataset. Further, we analyze the performance of the model when object annotation is manually given. Finally, we also discuss an application of the proposed visual attention model to identify the "best image" from a set of 2D images representing the light field image of a scene by defining a focus based visual attention metric. We also present a new eye tracking dataset on images captured using a light field camera. This dataset provides insight to the manner in which human visual attention is dependent on region of focus in an image and image semantics. This dataset was collected on 250 images from 21 subjects per image.

This work is organized as follows. In Section 6.1 we introduce the problem and review the related literature. In Section 6.2 we introduce the eye tracking dataset on images with varying camera focus. This is followed by the analysis of the eye movement regions in Section 6.3 In Section 6.4 we present our algorithm to learn visual attention maps by utilizing scene context features. In the subsequent Section 6.5 discusses the results of the visual attention maps learnt in our dataset and compares it to state-of-the-art. Finally, the summary and future work are discussed in Section 6.6.

## 6.1 Introduction and related work

There has been significant progress in light field camera technology in the past few years. This has resulted in wide availability of commercial light field cameras and it is an attractive alternative to traditional 2D imaging. Light field cameras are able to capture the intensity values for each ray direction thereby implicitly represents the 3D scene geometry. The growing popularity of light field cameras has led to several recent research efforts tailored for light field images in object recognition [65], depth estimation [99], segmentation [100], video acquisition [89], denoising and super-resolution [69]. Several such algorithms can greatly benefit from a localizing technique which identifies interesting regions by mimicking the manner in which humans process large streams of visual data. This visual attention model will make the algorithms robust and also enable them to allocate resources efficiently. Therefore, in this work we introduce a novel eye tracking dataset collected from multiple 2D images of a light field image captured using Lytro camera [38]. The camera focus region varies across the 2D images and this causes a significant change in the manner in which humans view the images. The example in Figure 6.1 clearly highlights the importance of camera focus in visual attention modeling. We notice a dramatic shift in attention when the foreground text is brought in focus compared to the background person.

Understanding the manner in which humans process visual stimuli is an interesting problem [17, 81]. Several research efforts in computer vision [66, 37, 56, 6, 20, 87,

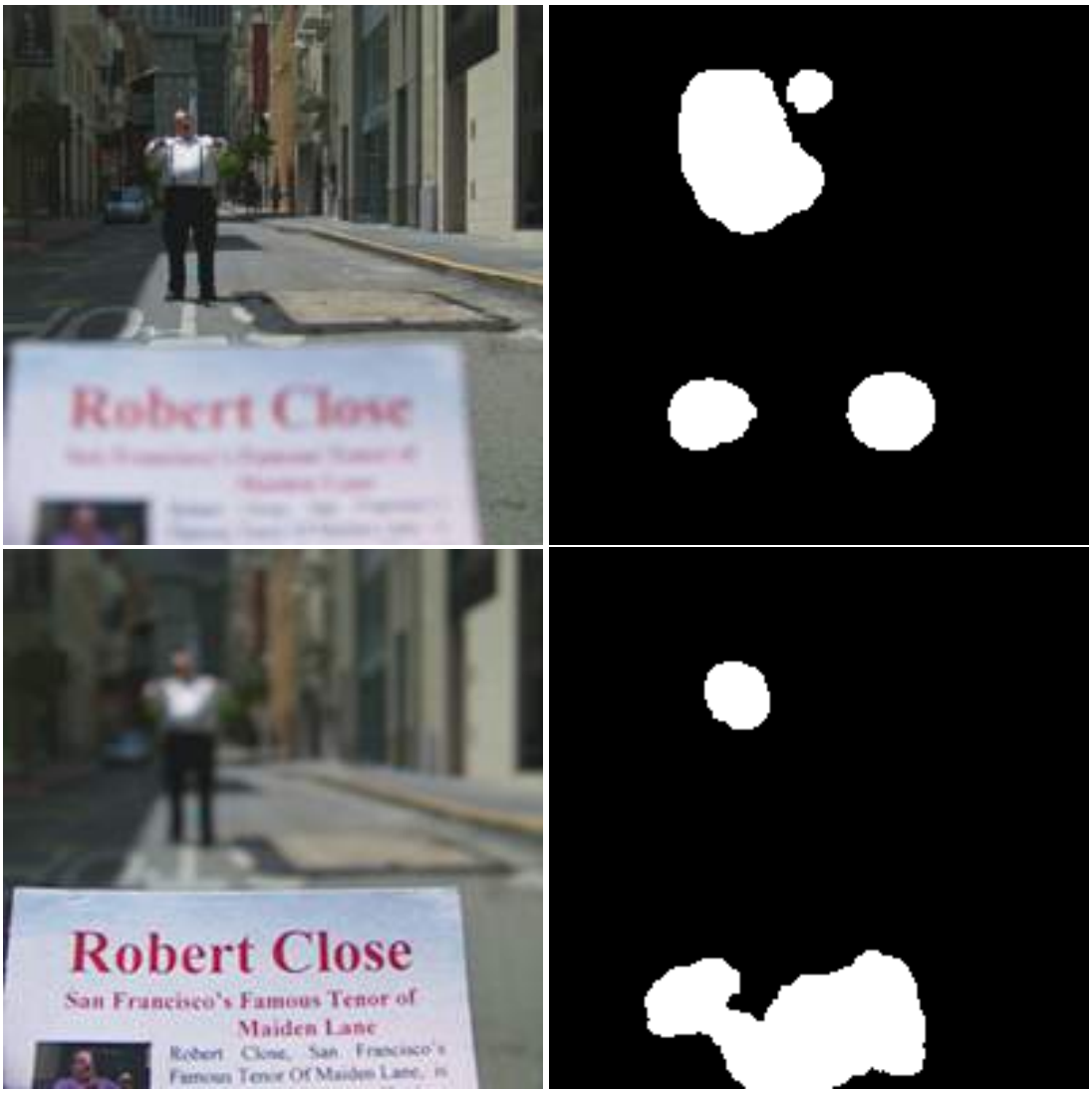



Figure 6.1: Left-Right. Top-Bottom (1) Image with background person in focus (2) Top 10% attention regions in the image from eye tracking data (3) Image with foreground text in focus (4) Top 10% attention regions in (3). We notice a significant shift in attention across two images of the same scene

97, 36, 95, 80, 77, 70], graphics [53], multimedia [63, 105], video compression [41, 50, 28] and robotics [83] have shown improved performance using visual attention models. Early visual attention models [51, 78] were pure bottom-up approaches. They

used several low-level image features such as color, texture and orientation in order to determine regions of interest in natural images. However top-down factors such a faces and text primarily attract visual attention [17]. In addition, image objects have shown to be better predictors of human fixations compared to bottom up saliency [30]. A model which utilizes this information to obtain improved human attention maps was proposed in [53]. The authors combine object detectors for car, person and face with low-level saliency maps and learn an support vector machine (SVM) [26] to predict human attention regions. The primary limitation of [53] is that it neglects object co-occurrence and scene context based analysis.

Recent research using controlled experiments [45, 64] highlight the importance of object co-occurrence and context for visual search tasks. These works indicate that other objects in a scene can provide a distracting(sometimes positive) effect for visual search of a specific object using reaction time studies. In addition context utilizing object co-occurrence plays a critical role in object recognition [72]. In a similar perspective, modeling object co-occurrence for a free viewing task helps in creating a better organization of interesting regions in a scene and our recent work in [56] introduces scene context features and uses a regression algorithm to identify interesting regions in images. The proposed visual attention model is inspired by [56] with additional scene context information relevant to camera focus. A detailed overview of various saliency algorithms and its applications are presented in [9] and a comparison of state-of-the-art

saliency benchmarks is presented in [10].

It is known that the efficacy of higher level contextual features to predict attention regions in images significantly depends on the reliability of object detectors. We have witnessed significant progress in object detection algorithms [35, 98, 58], however barring some non-deformable categories, reliability is poor for several object classes such as cats and dogs. In order to understand the gains of ideal object detectors, the proposed algorithm is also tested on human annotations of important objects in the scene. The primary contributions of this work are

- A model to predict attention regions in these images by utilizing low, mid, high and scene context features based on camera focus
- Evaluating the performance of our model using human annotations simulating ideal object detection scenario
- Predicting the best 2D still image from a light field image of a scene using a focus based attention metric
- A new eye tracking dataset on light field images captured using Lytro camera

## 6.2 Eye Tracking Dataset on Light Field images

We collected an eye tracking dataset, with Lytro images using Eyelink 1000 eye tracking device [3]. The dataset consists of 250 2D images from 105 Lytro images and from each light field image multiple 2D images were obtained by focusing at different depths corresponding to various objects in the scene. About 45% of the images were obtained from publicly available database [5]. We captured the ramining images our Lytro camera to get adequate images of both indoor and outdoor scenes. The dataset has good representation of common objects in natural images such as faces, person, text, car, dogs and cats. Figure 6.3 highlights some example images from the dataset. The images were re-sized to $1024 \times 768$ pixels which is required by the eye tracker and viewed by 21 subjects (between ages 16 and 35). The viewers sat 2.5 feet from a 27 inch screen and each image are shown for 3 seconds followed by 1 second viewing the gray screen. The subjects are instructed that it was a free viewing experiment and observe regions in images gather their interest without any prior bias. Also, eye tracking calibration is performed every 50 images (randomized order for each subject) and the entire data is collected in two sessions (125 images each).

Humans eye movement scanpaths typically consists of alternating fixations and saccades. Fixations represent information gathering sequence around an interest region and saccades represent transitions between fixations. The eye tracking host computer samples the gaze information at 1000 Hz and automatically detects fixations and sac-

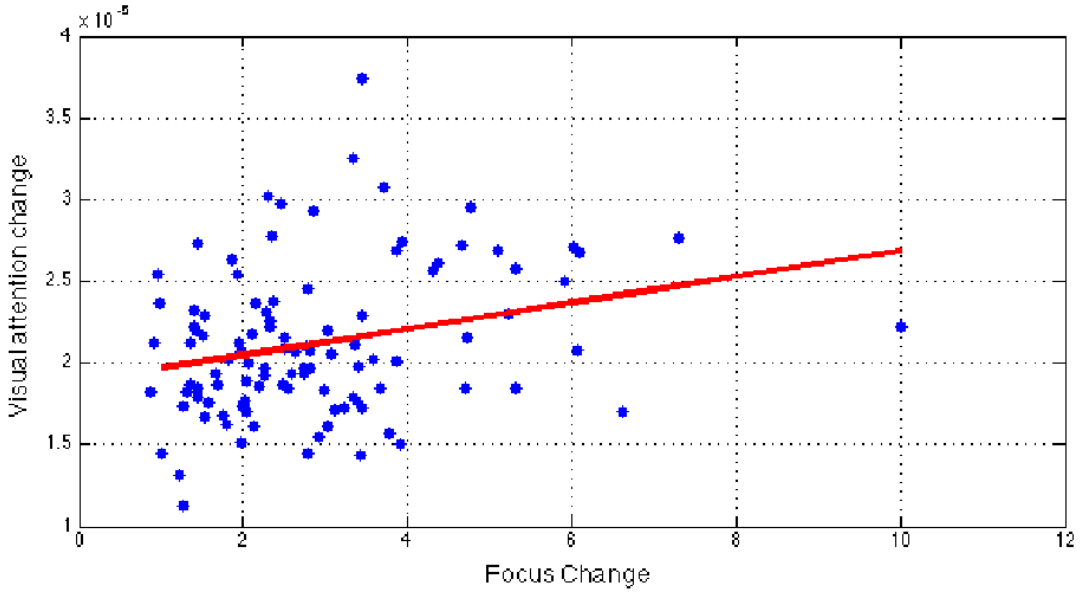



Figure 6.2: Plot of focus change with attention change. We notice that there is noticeable correlation (shown by red line) between visual attention change and change in camera focus. The correlation is represented by the best line fir to this data which has a slope of 0.27.

cades in the data. The eye tracking device also clusters the fixation samples and identifies fixation and saccade points. We use these fixation and saccade points to create the ground truth visual attention maps. The number of fixations and saccade points per image can vary from 6 to 15. In our experiments, the first fixation and saccade is removed to avoid the initial eye position bias directed by the transition gray slide in the experimental setup. The final human visual attention ground truth map is obtained by filtering the fixations using a Gaussian filter.

## 6.3 Analysis of eye tracking dataset

In this section we aim to understand the properties of eye movements across camera focus. We identify in-focus and out-of-focus regions in an image using an approach similar to [34]. Let $L$ be a Lytro image and $\{I_i\}$ represent the 2D images from $L$. For image $I_i$, the gradient image $G_i$ is computed and is divided into non-overlapping sub-windows (we used $10 \times 10$) and the image gradient magnitude ($G_i^{abs}$) and standard deviation ($G_i^{std}$) is computed on these sub-windows. The measure of focus in each image pixel is square root of the product of the gradient magnitude and standard deviation as shown below.

$$F_i(p) = \sqrt{G_i^{mag}(p) G_i^{std}(p)} \forall p \in \text{pixels} \tag{6.1}$$

In some applications we require a binary threshold of whether the region is in-focus or out-of-focus. This can be easily accomplished by thresholding the focus map $F_i(p)$ and create a binary focus map. Figure 6.4 shows some example binary focus maps generated using the proposed algorithm.

### 6.3.1 Visual attention variability with camera focus

The first study understands the importance of how change in camera focus affects visual attention. For each lytro image $L_i$, we calculate the average focus change by

computing the focus $F_j$ for each image $I_j, j \in 1...M$ using Equation 6.1. If there are $M$ 2D images representing a lytro image, there are $\frac{M(M-1)}{2}$ unique image pairs where we measure focus change and the average focus change metric $FC_i$ for $L_i$ is computed as

$$FC_i = \frac{\sum_{j=1}^{M-1} \sum_{k=j+1}^{M} 2||F(i) - F(j)||^2}{M(M-1)} \tag{6.2}$$

The visual attention change from the ground truth masks is computed in a similar manner. Let $G_i$ represent the ground truth visual attention map for an image. The average visual attention change metric $VC_i$ for $L_i$ is computed as follows

$$VC_i = \frac{\sum_{j=1}^{M-1} \sum_{k=j+1}^{M} 2||G(i) - G(j)||^2}{M(M-1)} \tag{6.3}$$

The focus change ($FC$) is plotted against visual attention change ($VC$) in Figure 6.2 and we notice a correlation of 0.27 between focus change and visual attention change which is significant. This analysis shows that standalone camera focus alters visual attention and higher change in focus has some correlation to greater attention change. In the following analysis we go beyond focus based analysis by accounting for objects in the scene and understand how in-focus and out-of-focus objects affect visual attention.

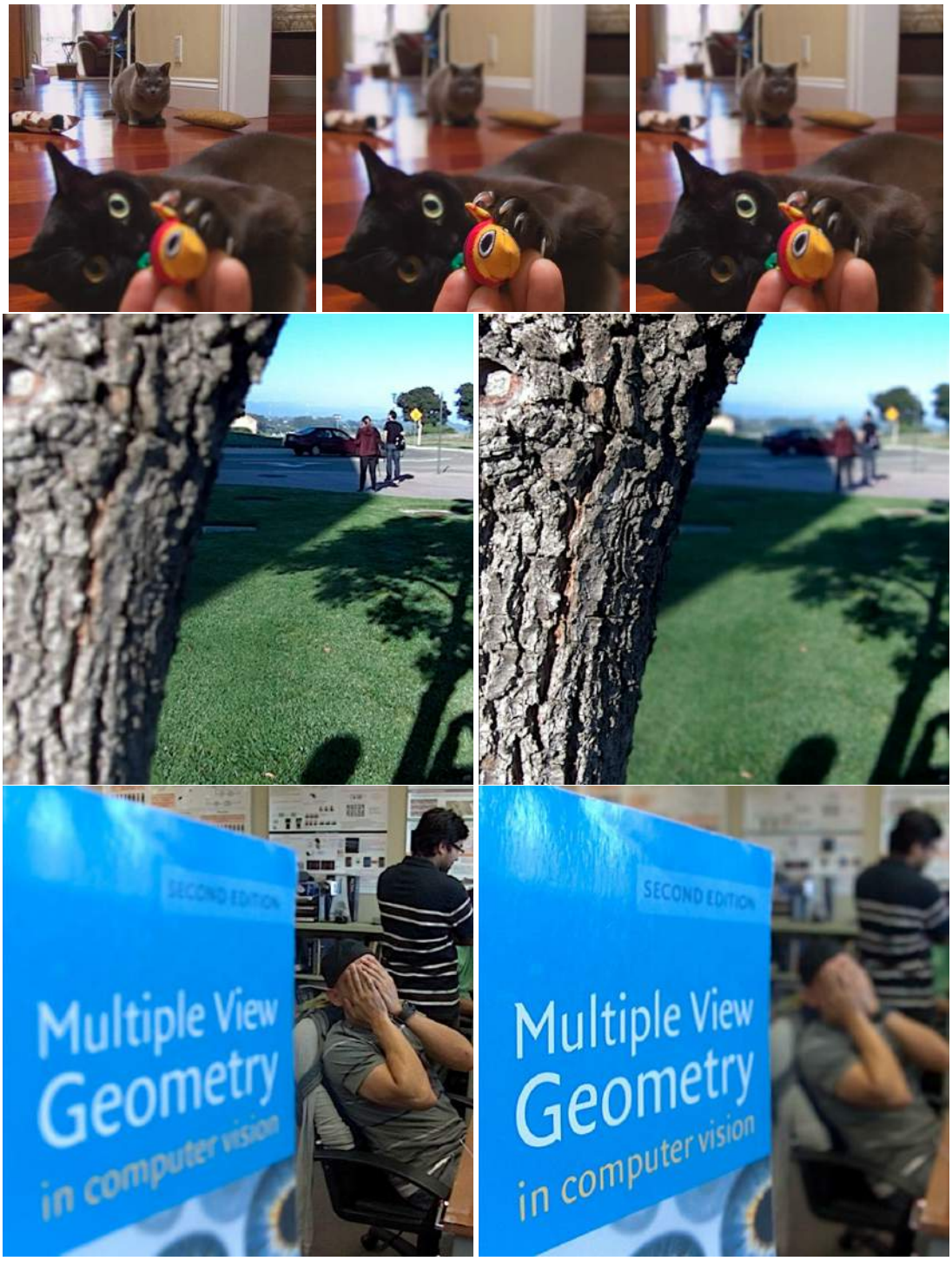



Figure 6.3: Examples of images from our dataset consisting of faces, text, people, animals and car etc. at different focus depths.

Figure 6.4: Two 2D images from a Lytro image and the corresponding binary focus maps to their right.

### 6.3.2 Object oriented focus based attention analysis

The dataset also includes manual annotations of important objects in each scene. In total we have annotated 16 different objects in the scene. Some commonly occurring categories are further divided according to their size as small or large as visual attention properties depend significantly on the size of these objects. In total we have 23 different annotated object types and a few examples are shown in Figure 6.5. They include face, eyes, nose, mouth, person, text, animal, toy, flower, vehicle, electronics, plant and building and categories such as face, text, animal, toy, flower, vehicle and text are divided into small and large as they commonly occur in a variety of sizes. First, we find that about 86% of all the visual attention maps obtained from fixations fall in the regions annotated as objects. This highlights the primary role that high level semantics play in modeling human visual attention. Also, the binary focus masks computed according to Equation 6.1 are used to categorize in-focus and out-of-focus objects based on simple majority vote. The visual attention density for each category is calculated and shown in Figure 6.6 over focused and out-of-focus objects. We notice that camera focus plays a critical role in determining the attention density captured by an object. Typically, small objects have higher attention density than large objects as the density is averaged over the number of pixels in the objects. Several important high attention density categories such as small faces, text and animals which have sufficient representation in our dataset experience about a two fold increase in visual attention when the objects are brought

in-focus. In addition, the corresponding large text and animals have about 20% increase in visual attention, however large faces have a more significant gain in visual attention when they are in focus, about 60%, compared to out-of-focus large faces. This analysis clearly portrays that camera focus is a significant factor in determining how humans perceive images.

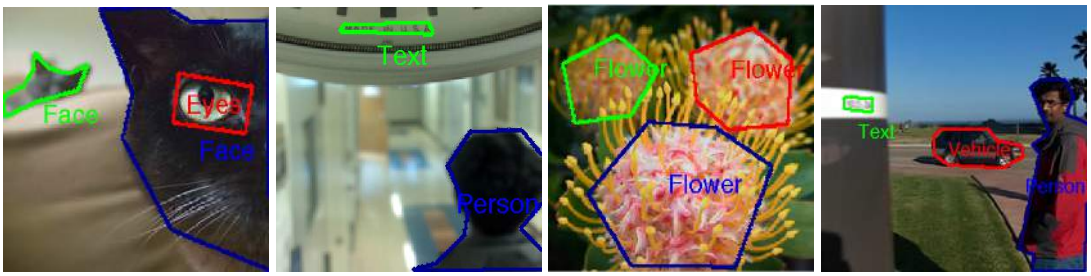



Figure 6.5: Examples manual object annotations

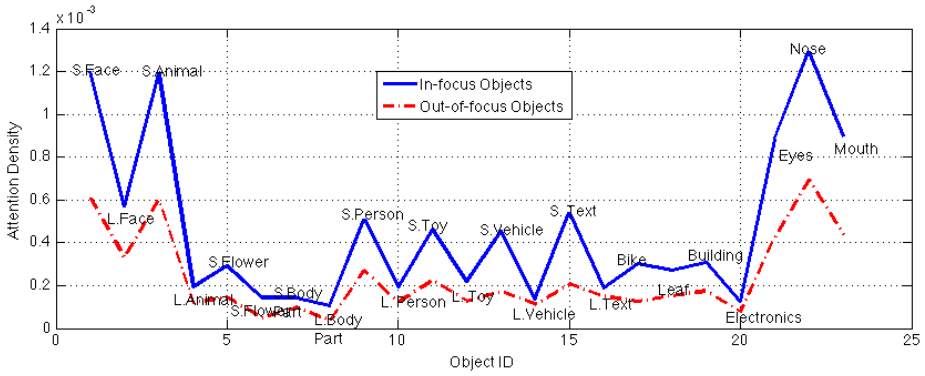



Figure 6.6: Attention density in 23 manually annotated objects in-focus (blue) and out-of-focus (green). Objects in focus consistently have higher attention density than out-of-focus objects

### 6.3.3 Initial fixation statistics over camera focus

The previous section analyzes visual attention maps obtained from all fixations are distributed across different object categories. In addition, it would be interesting to study the effect of camera focus on initial fixations when a subject observes an image. Typically the initial fixations are biased towards the most important concept or object in the scene, and therefore we expect the importance of an object in a scene to be enhanced when they are in focus and correspondingly expect earlier initial fixations compared to when they are out of focus. Figure 6.7 highlights this trend in our dataset, where the average number of fixations before the first fixation to an object is computed over all the images. First, we notice that larger objects are fixated earlier than their smaller counterparts. For the most commonly occurring objects such as faces, text and animals in our dataset, on an average we observe the first fixation is delayed by about 2 fixations for the small objects compared to the corresponding large objects. In addition, we observe focus plays a significant role in determining the average fixation number of the first fixation consistently for all objects. Especially for small objects, we notice around 1.5 fixations delay when out-of-focus. However, in large objects the first fixation is not significantly affected by camera focus for several categories with faces being an exception. Previous studies [81, 17] have shown that initial fixations are biased towards faces, however results in our dataset show the time to first fixation to faces is not considerably lower than other categories. This can be attributed to the

fact that several images in our dataset consist of multiple face images and though the initial fixations are biased towards one of the faces, the average time for the first fixation to any face is high due to the presence of other faces/people. This leads to increased average number of fixations before the first fixation to a face.

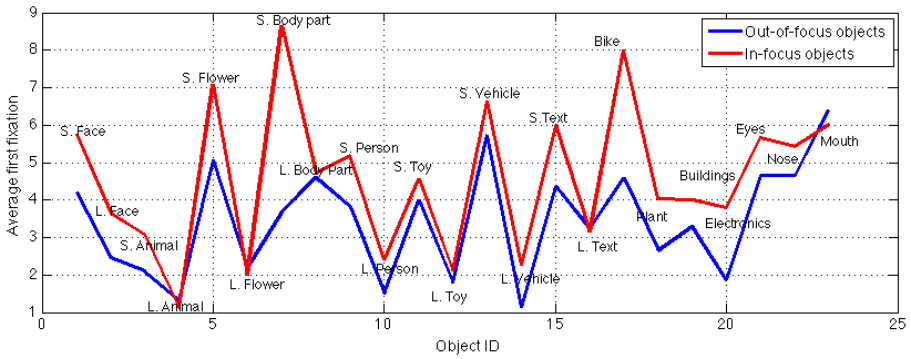



Figure 6.7: Average initial fixation time in 23 manually annotated objects in-focus (blue) and out-of-focus (red). Objects in focus consistently have lower average time to first fixation compared to out-of-focus objects

### 6.3.4 Visual attention consistency across camera focus

The previous sections highlighted the manner in which camera focus changes the way in which we observe objects in natural scenes. However, there is considerable consistency as well in the way in which observers look at images of the same scene which only differ by region of focus. The performance of visual attention consistency is presented using ROC curves, which are computed as follows. For each image we have a ground truth map and a predicted saliency map from an image from differ-

ent focus computed as described in Section 6.4. This saliency map is thresholded at k=1,3,5,10,15,20,25 and 30 percent to obtain binary saliency maps. The percentage of ground truth map contained within each binary saliency map is the performance measure of how attention maps from different focus predicts the ground truth. It is well known that average fixations of other subjects are a good predictor of a new observer's fixations. Therefore, the consistency of fixations across multiple foci is compared to the consistency of fixations among multiple subjects in the same image. This curve which measures consistency of fixations among multiple subjects is called the human ROC curve. Figure 6.8 compares the performance of the human ROC curve to focus ROC curve and we notice a drop of about 10% in the focus ROC curve.

We also highlight examples of cases where visual attention is consistent across camera focus and cases in which focus drastically changes visual attention in Figure 6.9. The plots of the individual ROC curves of the examples shown in Figure 6.9 are shown in Figure 6.10. A significant drop and gain in visual attention consistency in the respective images is noticed compared to the mean scenario in these examples.

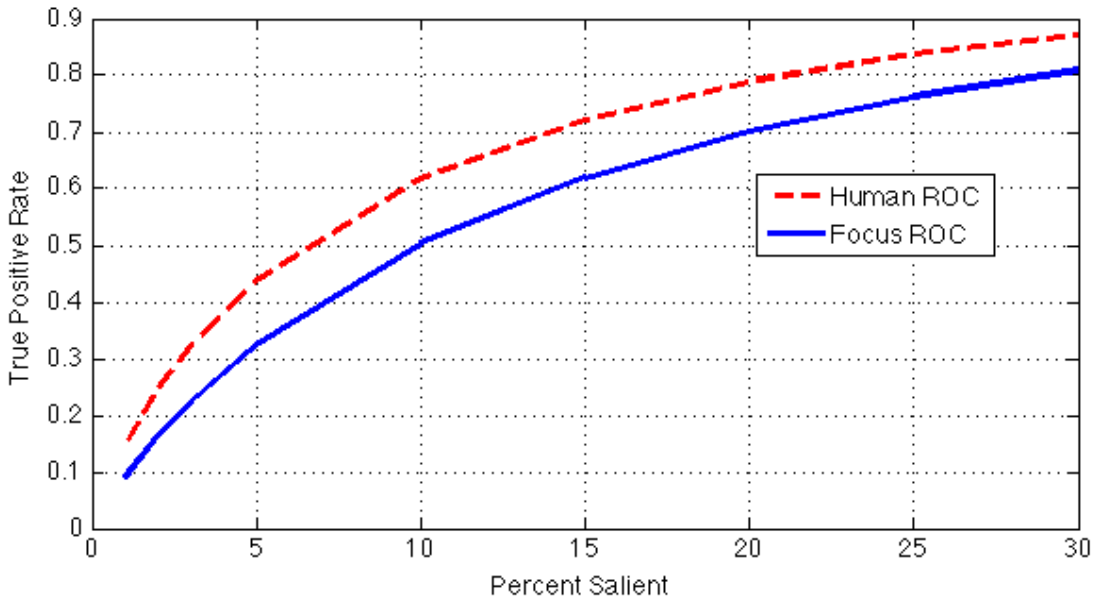



Figure 6.8: Visual attention consistency across multiple focus images (Focus ROC curve) . It is shown in reference to the human ROC curve which measures the consistency among fixations in the same image across multiple subjects.

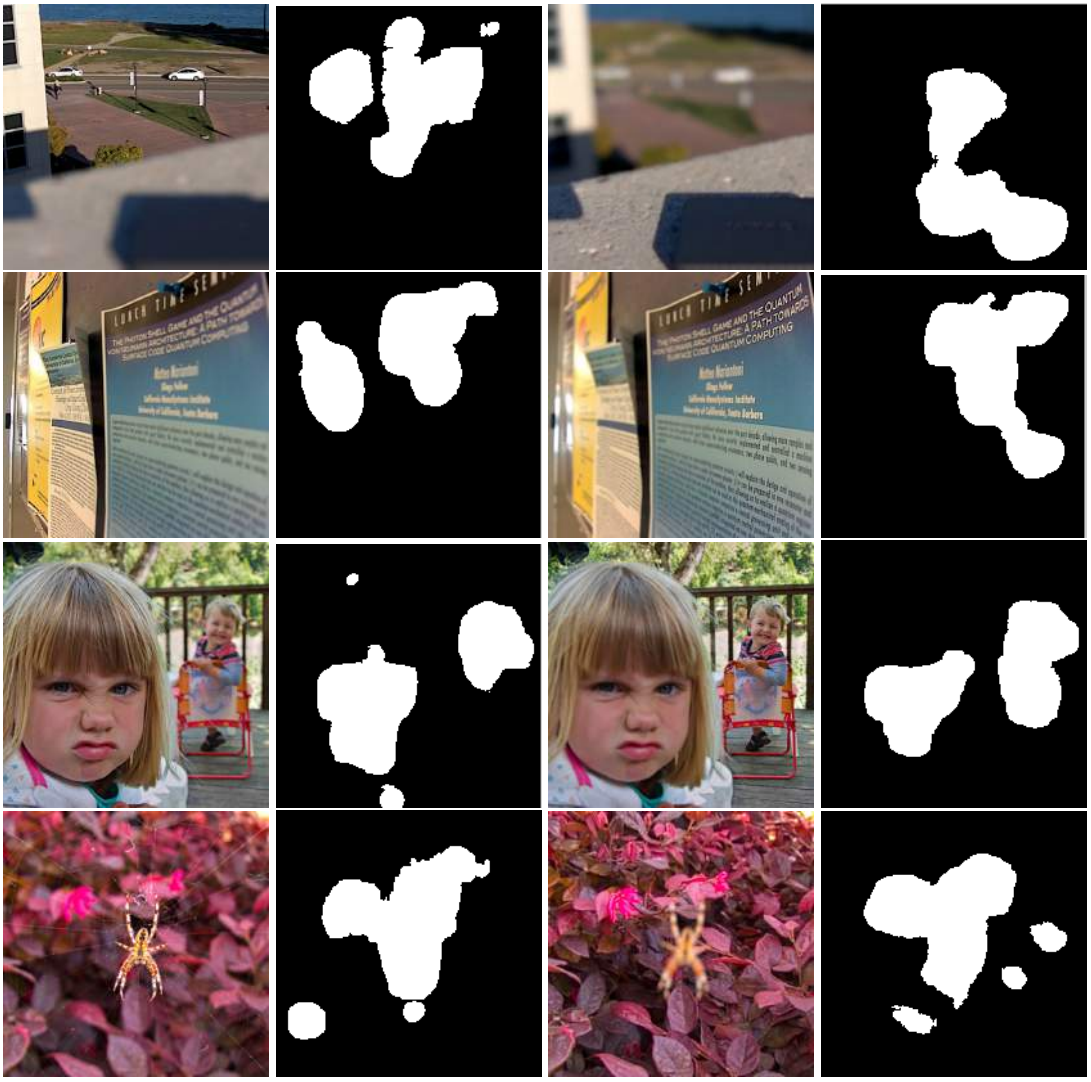


Figure 6.9: Row 1 and 2 shows examples of cases where camera focus significantly alters visual attention. We notice categories such as text can have a significant change in the manner in which we view images if their focus attributes change. Rows 3 and 4 highlights examples where camera focus does not significantly alter visual attention. In row 3, the out of focus faces also attract significant attention. In row 4, there is only one salient object in the center of the image and therefore attracts considerable attention irrespective of camera focus.

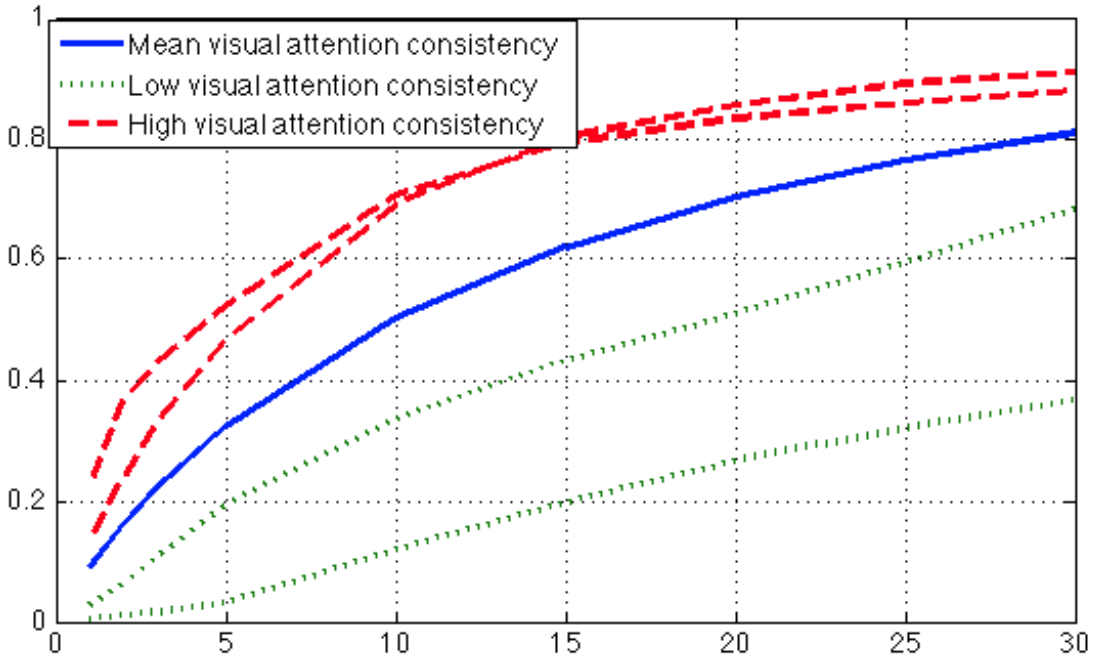



Figure 6.10: The plot shows the consistency of visual attention in the images shown in Fig 6.9

## 6.4 Visual attention model for light field images

In this section, we propose an algorithm to predict salient regions in an image using image features. Our saliency algorithm learns a regression model from features extracted at multiple levels in an image. In addition to the low, mid and high level features used in [53], the proposed technique also uses scene context features which model the interaction between these features. The following sections explain the proposed feature extraction and learning steps.

### 6.4.1 Feature extraction

#### Low level features

Our model utilizes the following low level features due to their importance in bottom up saliency.

Itti and Koch saliency: Early saliency model provides intensity, color and orientation channels [51] motivated by linear filtering and center surround operation, suitable for bottom up visual attention modeling.

Steerable pyramid filters: These filter responses correlate well with visual attention and therefore local energy of steerable pyramid filters [84] in three scales and four orientations are used.

Torralba Saliency: Provides a holistic representation of a scene [71] using coarsely lo-

calized spectral information.

Signature Saliency: Provides a saliency map [46] which spatially approximates image foreground using the theoretical framework of sparse signal mixing.

Graph Based Visual Saliency: Jointly models activation map creation and feature extraction in a unified manner by defining edge weights using dissimilarity and saliency [43].

**Mid level features**

Horizon detection is performed using using gist descriptor [71]. It is especially important in outdoor scenes where salient objects are present near the ground plane .

**High level features**

High level objects such as faces and text have high visual saliency. We utilize automatic object detectors for face [93], person [35], car [35] and text [104] in our model. Also, from our analysis in Section 6.3, we notice that camera focus plays a major role in determining visual attention for high level objects as shown in Figure 6.9. We notice that objects in focus consistently have higher attention density than out-of-focus objects according to Figure 6.6. Therefore, the focus regions in an image are computed according to Equation 6.1 and each detected object is identified as in-focus or out-of-focus, based on whether a majority of the pixels are in or out-of-focus

respectively. Therefore the number of high-level features is doubled from 4 to 8 in the proposed algorithm.

**Scene Context features**

In addition to high level features, we utilize features which model the pairwise interaction between multiple high level features inspired by our previous work in [56]. High level features typically model attention gain in the locality of semantic objects. However, presence of interesting objects in a scene also incurs attention loss in other objects in a scene. This attention loss scene context features can be described using a *cause-effect* mechanism. Let there be $N$ possible objects in a scene and $f_{ij}^{SC}(x,y)$ denote the scene context feature between object $i$ and $j$ at position$(x,y)$. The scene context vector models the attention loss in the scene incurred on the pixels of object $i$ (*effect*) due to presence of object $j$ (*cause*). Now let the total number of objects corresponding to label $i$ in the scene be denoted by $N_i$ and the number of object $i's$ in position $(x,y)$ be $n_i(x,y)$. Let the image be denoted as $\mathcal{I}$, the scene context vectors are defined as

$$f_{ij}^{SC}(x,y) = \begin{cases} N_j - n_j(x,y) & \text{if } n_i(x,y) > 0 \\ 0 & \text{if } n_i(x,y) = 0 \end{cases} \tag{6.4}$$

Figure 6.11 presents an example of scene context features where $N = 3$. Object 3 is not present in the image and therefore $f_{ij}^{SC} = 0$ if $i = 3$ or $j = 3$. Next, $f_{11}^{SC}$ informs

the regions where object 1 is present about another object 1 detection. Further, $f_{21}^{SC}$ indicates the pixels of object 2 the presence of two object 1 detections.

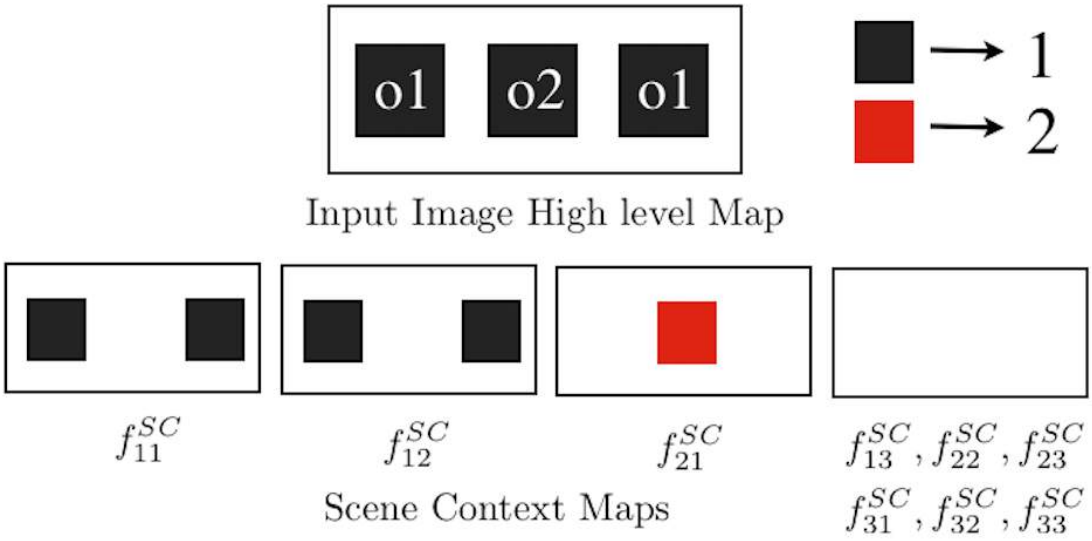
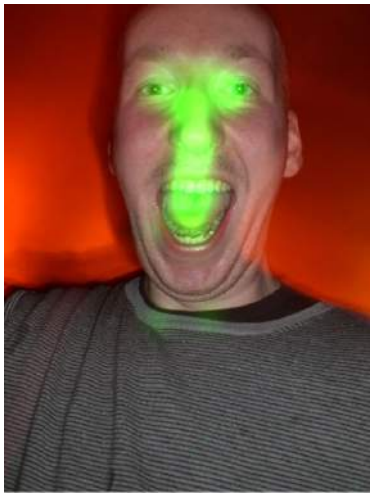



Figure 6.11: An example high level object layout with N=3 and the corresponding 9 scene context feature maps

Typically, the number of distinct objects which can occur in an image is large and modeling the cause-effect relationship between every pair of objects is necessary. However, it is not possible to obtain such large number of pairwise interactions in practical eye tracking datasets. Therefore, to reduce the dimensionality of the scene context features, we utilize an approximation technique proposed in [56] which clusters the cause and effect features separately that have similar properties.

*Cause effect clustering:* Our aim is to model the factors affecting cause and effect of per pixel attention loss (attention density loss) which will be predicted by the learning

(a) Car



(b) Face

(c) Car and Face

Figure 6.12: (a) Has only one salient object (small car) and (b) has one salient object (large). However (c) has both the salient objects. We observe the large face significantly draws attention (green overlay) away from the small car in (c). Best viewed in color.

algorithm. In order to gain insight to factors affecting cause and effect of attention loss, consider an example in Figure 6.12 which shows individual images of a small car (left), large face (center) and an image where they co-occur (right) with the corresponding attention map overlays in green. The attention sum of the two large faces in (b) and (c) are 0.7 and 0.6 and densities (normalized) are 6 and 5 respectively. The attention sum of the two cars in (a) and (c) are 0.1 and 0.05 and densities (normalized) are 20 and 8 respectively. The attention density in the large face remains relatively unchanged due to the presence of the small car, but the presence of the large face significantly contributes to attention loss in the small car as its density is almost halved. This illustrates the requirement to cluster the attention loss cause and effect features with different metrics as presented in Algorithm 4. The algorithm describes the computation of the attention sum ($sA$) and density ($dA$) for all high level objects. The cause clusters ($\mathcal{C}$) and effect clusters ($\mathcal{E}$) are derived from the attention sum and density features respectively. Objects which cause a large attention shift are typically objects which have a large overall attention sum contained within them(typically objects which occupy a large portion of viewing area, eg. large face in Figure 6.12). Also, the objects which take the greatest impact or effect have high concentration or attention density (small prominent objects, eg. small car in Figure 6.12).

**Data**: $\{\mathcal{I}_i\}$ - Input Training Images. $\{\mathcal{A}_i\}$ - Attention maps.

$\{\mathcal{H}_i^{jk}\}$ - Binary high level maps denoting the $k^{th}$ occurrence of high level object index $j$ in image index $i$

$N-$number of training images, $M-$number of high-level objects

$i \in [1, N]$ and $j \in [1, M]$

**Result**: $\mathcal{C}$ - Cause Clusters, $\mathcal{E}$ - Effect Clusters.

**Initialization:** $n^j = 0 \forall j \in [1, M]$ (counter)

$dA = \emptyset$ (Attention density vectors) $sA = \emptyset$ (Attention sum vectors);

**for** *i=1*$\rightarrow N$ **do**

  **for** *j=1*$\rightarrow M$ **do**

    **for** *k=1*$\rightarrow |\mathcal{H}_i^{j*}|$ **do**

      $n^j = n^j + 1$;

      $sA(n^j) = \sum_{p \in \mathcal{P}} \mathcal{A}_i(p) \bigodot \mathcal{H}_i^{jk}(p)$;

      $dA(n^j) = \frac{\sum_{p \in \mathcal{P}} \mathcal{A}_i(p) \bigodot \mathcal{H}_i^{jk}(p)}{\sum_{p \in \mathcal{P}} \mathcal{H}_i^{jk}(p)}$;

    **end**

  **end**

**end**

$\mathcal{C}$ = kmeans($sA$); $\mathcal{E}$ = kmeans($dA$);

**Algorithm 4:** Training algorithm

#### Center Prior

Typically images contain the object of interest in the center. Therefore, we utilize a center prior map [53].

### 6.4.2 Learning

The features are pre-computed in all the images and a partial least squares (PLS) regression [101] model is used to predict where subjects look in new images. The PLS regression algorithm is selected as it obtained better prediction results than other techniques such as linear regression [24], random forest regression [12] and lib-linear SVM [33] in [56] for visual attention modeling. The entire dataset is divided into training and test sets in a 15-fold cross validation setting. From each image in the training dataset, we randomly pick equal number of pixels from the top 20% and bottom 80% attention regions (to have adequate representation for high attention regions) to create a pixel level training subset. A PLS-regression model is learnt from this subset and pixel wise attention density in new images are predicted using this regression model.

## 6.5 Experiments and Results

### 6.5.1 Setup

For every image, low level saliency features and high level object detections are computed as described in Section 6.4.1. Object size plays a significant role in determining attention density per pixel. Therefore, objects are distinguished as small or large by a size threshold (2500 pixels in our attention maps). Hence, the number of distinct objects double from 8 to 16. Further, we obtain 9 scene context features similar to the procedure described in [56]. To sum up, *we have 30 low-level features, 1 mid-level feature, center prior, 16 high level features and 9 scene context features totaling 57 features.* These features are used to obtain the predicted saliency map using PLS regression algorithm as described in 6.4.2.

### 6.5.2 Visual Attention modeling

We present a comparison of the performance of various saliency algorithms with the proposed approach. Performance of saliency algorithms is presented using ROC curves, which are computed as follows. All visual attention algorithms generate a saliency map with a predicted pixel-level saliency. This saliency map is thresholded at $k$ =1,3,5,10,15,20,25 and 30 percent to obtain binary saliency maps. The percentage of human fixations contained within each binary map is the performance measure.

A comparison of performance of the proposed approach using PLS regression [101] with Judd et al. [53], GBVS [43], Itti and Koch saliency [51], Torralba saliency [71], Signature saliency [46] , Center prior (DistCenter), SUN Saliency [107] and chance is shown in Figure 6.13. Our algorithm outperforms other saliency algorithms by a significant margin. Judd et al. [53] also utilizes high level objects (car, person and face) but does not model focus based object attention and learns only a single weight for every object irrespective of the size of the object and scene statistics. Hence, on an average the proposed approach obtains 3.4% gain over [53]. Figure 6.14 shows some examples of light field images with the corresponding visual attention maps from the proposed approach. We notice that the proposed approach is a good predictor of ground truth visual attention and has the inherent ability to transform the predicted visual attention when camera focus changes. Among algorithms which do not employ object detectors, we notice that GBVS gives the best performance followed by signature saliency. We additionally evaluated the performance of AIM [13] and Spectral Residual [47] saliency algorithms but their performance was lower than the compared approaches in Figure 6.13.

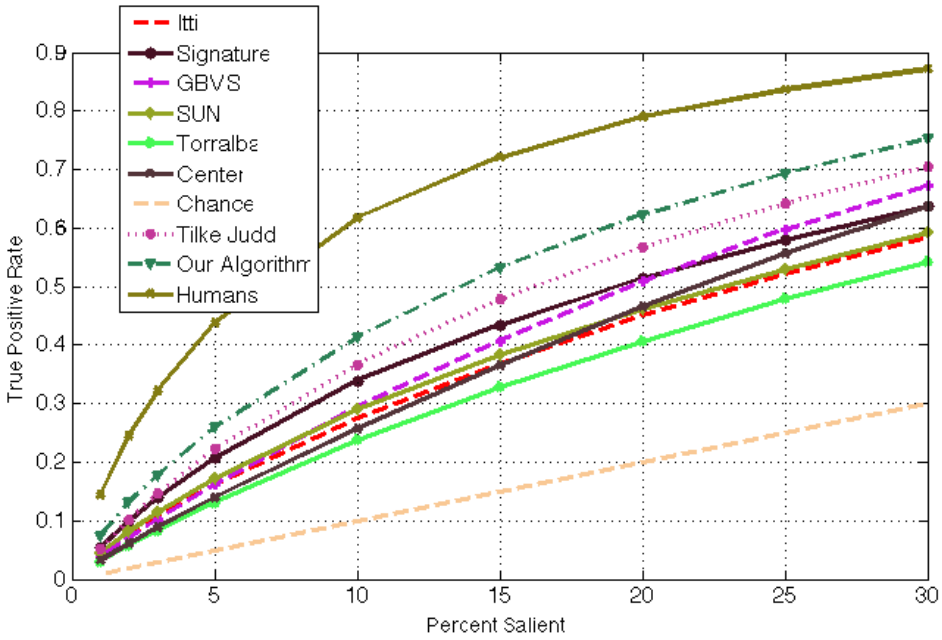



Figure 6.13: Comparison of the performance of the proposed approach with other state-of-the-art visual attention and saliency algorithms using ROC curves. Our approach outerforms other algorithms by a significant margin. However, we still notice considerable gap between machine and human performance.


Figure 6.14: Each row contains ground truth attention map (Blue overlay) with a focus setting followed by corresponding predicted visual attention map (Red overlay). The results are shown on two Lytro images (rows) each under two focus settings

### 6.5.3 Ideal Object Detection Scenario

From the previous section we notice that object detectors play a major role in improving visual attention prediction. In addition in Section 6.3.2 we observed that a significant portion of fixations were present in important objects in a scene. Therefore, we are interested in evaluating the upper bound performance when object detectors are perfect using our proposed model. We have annotated thirteen different object categories face, text, animal, flower, building, etc as shown in Figure 6.5. In total, we have 23 high-level size based features and we have 46 focus based object features. In addition we have 16 scene context features (4 cause and effect clusters each, gives 4×4 scene context features) as well. In total we have 98 features including low, mid and high-level features. A comparison of the performance of annotated objects with automated object detectors is shown in Figure 6.17. We notice that we get about 8% improvement in performance using annotated objects on an average, however in order to bridge the gap with human performance, more research on higher level semantic context is necessary.

The regression weights of different objects, which highlight the importance of each feature to predict the saliency of a pixel is shown in Figure 6.16. This comparison is feasible as we normalize each feature vector using the its mean $\mu$ and standard deviation $\sigma$ as $x \sim \frac{x-\mu}{\sigma}$ and bring all the features to a comparable common platform. We notice that the regression weights are reminiscent of the average visual attention density shown in Figure 6.6 and therefore the regression weights are indicative of the attention density

in an object.

In addition the scene context features weights are illustrated in Figure 6.15. These weights are expected to capture attention loss in an object due to the presence of other surrounding objects in the scene. We notice that compared to the average value of regression weight for non-scene context feature which typically has an additive effect, scene context feature weights are negative corroborating our intuition in designing the feature.

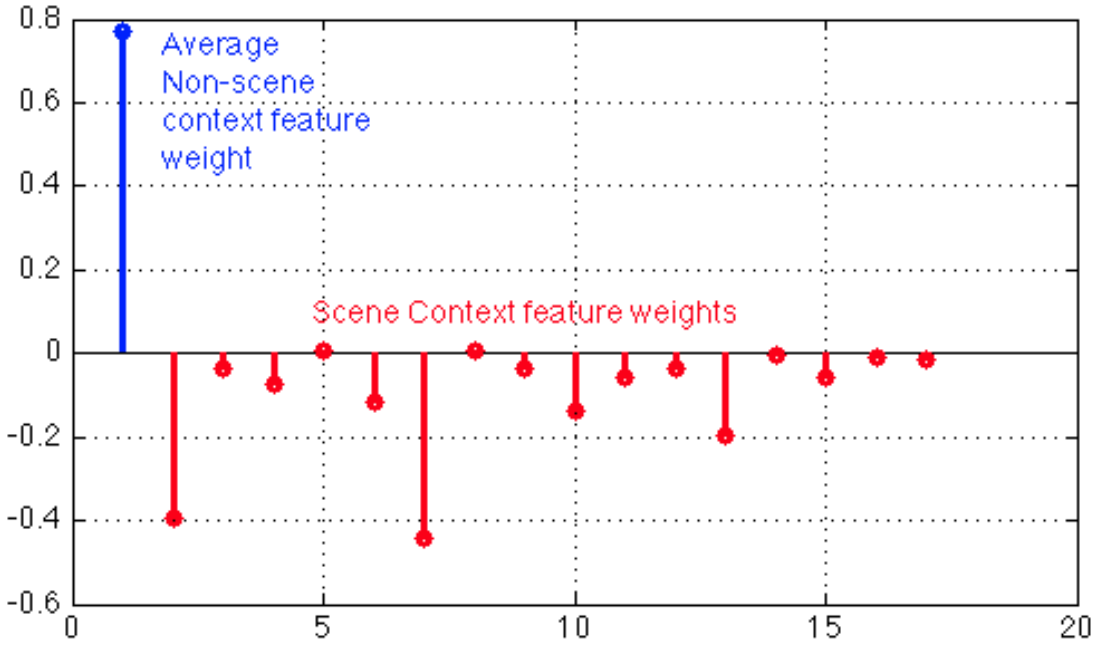



Figure 6.15: Weights of scene context features compared to average non-context feature

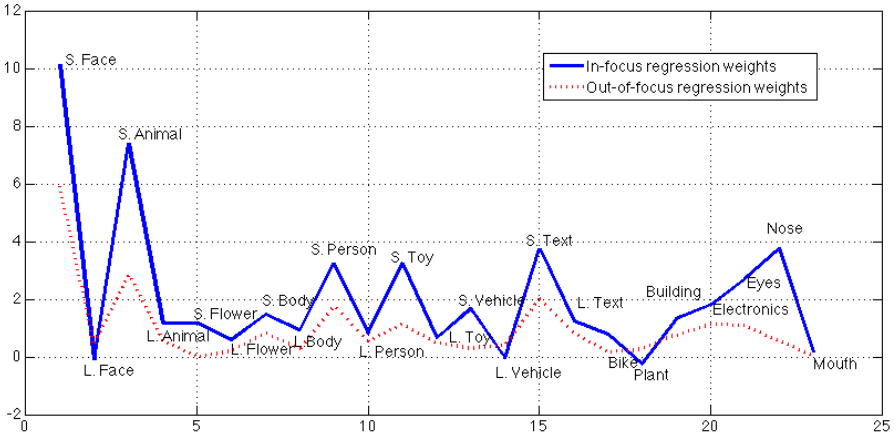



Figure 6.16: Regression weights of different objects in in-focus and out-of-focus object categories. We notice this plot is similar to the average attention density plot in Figure 6.6

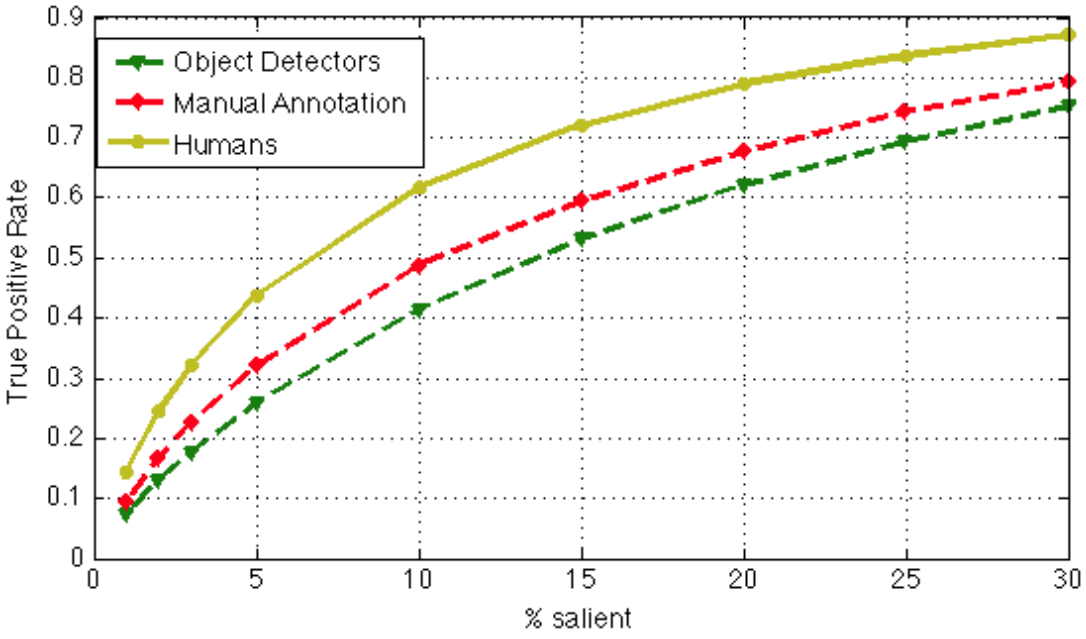



Figure 6.17: Performance of manual annotations compared to automated detectors.

| | Judd et al.[10] | Proposed approach | Random |
|---|---|---|---|
| Accuracy | 72.8 | 75.3 | 40.0 |

Table 6.1: Comparison of the performance of Judd et al. [10] with the proposed approach to predict the best 2D image from a Lytro image

### 6.5.4 Classifying best image using focus based attention

We present a practical application for the proposed visual attention model in this section. The primary rationale behind bringing a region in focus is to bring the viewer's attention to the scene situation. Therefore, given a set of 2D images from a light field image we define the best image as the one which has the largest attention concentrated in the region of focus. This approach can also be viewed as a mechanism for semantic auto-focus. An example to identify the best image on two Lytro images is shown in Figure 6.18. Formally, given a set of images $\{I_i\}$ from a Lytro image $L$, with the ground truth map attention map $G_i$, with binary focus map $F_i$ defined according to Equation 6.1, the best image is defined as follows

$$\text{Best-Image}(L) = \arg\max_i \sum_{p \in \text{pixels}} F_i(p) G_i(p) \tag{6.5}$$

The problem here is to predict the best image using the learnt visual attention maps and calculate the percentage of correct identification. Table 6.1 shows that the proposed approach identifies the best image with 75.2% accuracy and the best competing approach [53] in Table 6.1. identifies the best image 72.8% in the 105 Lytro images.

Figure 6.18 shows two examples where the best 2D image is indicated from a Lytro image based on ground truth visual attention in the focused region.

## 6.6 Summary and future directions

We introduce a new eye tracking dataset for light field images captured using Lytro camera. The eye tracking data was collected on 2D images by changing the focus regions and 2-4 distinct focus regions are observed for each Lytro image. We analyse the collected eye tracking data and observe that camera focus and object localization are significant factors in the manner in which humans view images. A visual attention model is learnt utilizing low, mid, high and scene context information with emphasis on in-focus and out-of-focus objects. The proposed model outperforms other state-of-the-art visual saliency models in our eye tracking dataset. We also analyze the performance of the proposed model using human annotated objects for thirteen categories. A significant gain in performance is noticed but we also realize the requirement of higher level contextual information to bridge the gap with human performance. Finally, a new problem of predicting the best 2D image from a Lytro image by defining a focus based visual attention metric is introduced. We notice our technique outperforms [53] to identify the best still image from a Lytro image.

The performance of current visual attention models have significantly improved by

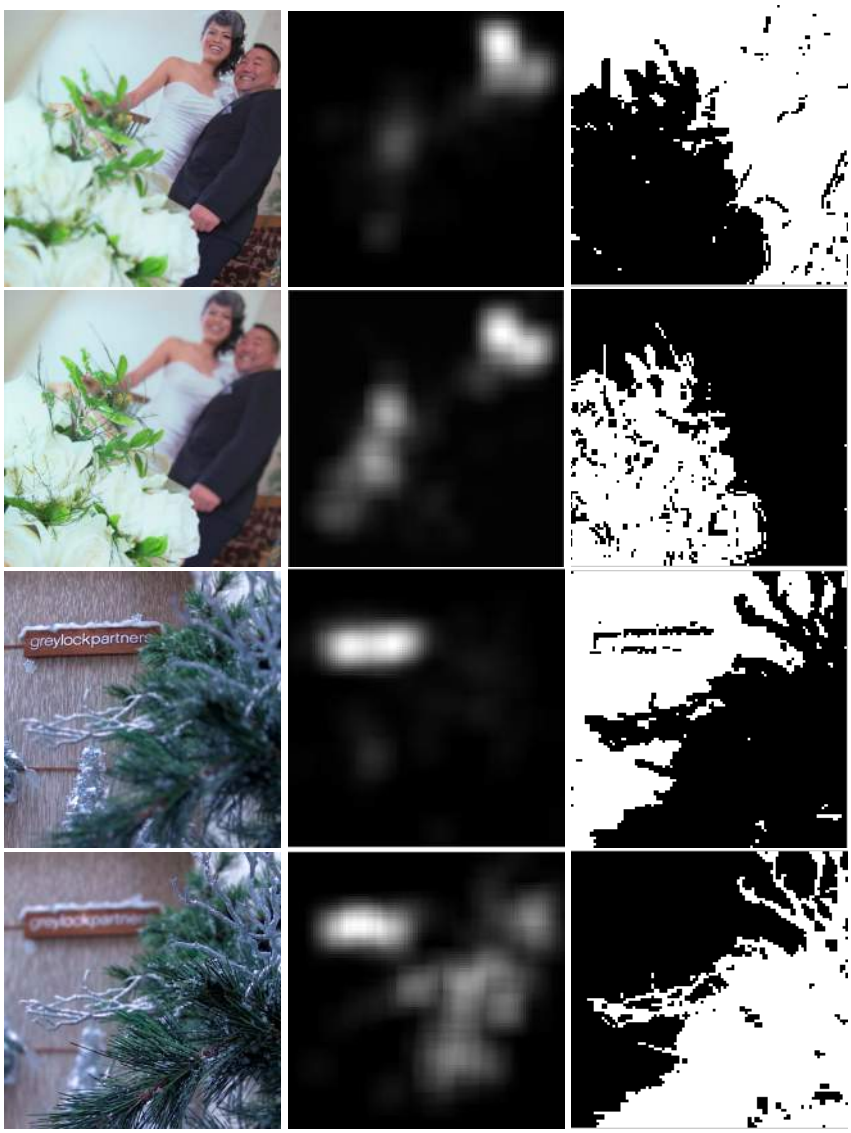


Figure 6.18: Left column indicates the input image. The center column indicates the visual attention map and the right column indicates the region in focus. An analysis of row 1 and 2 indicates that the image in row 1 has higher visual attention in the focused region than image in row 2. In addition we notice that image in row 3 has higher attention in the focused region than image in row 4. Therefore, input image in row 1 and row 3 are preferable to row 2 and row 4 respectively as they capture larger visual attention in the region of focus.

the utilization of semantic information from visual scenes. Therefore, advancement in computer vision algorithms will naturally improve the robustness of higher level semantic information mining from images which will improve the prediction of visual attention map. Additionally, we can explore different paradigms to tackle the visual attention prediction problem. The proposed visual attention maps require prediction of pixel level attention from image features. However, the visual attention map is generated from human fixation map from multiple subjects. Therefore, it would be interesting to explore more algorithms to directly predict the human fixation map, from high level semantics, instead of the smoothed visual attention map.

# Chapter 7

# Conclusions and Future Work

"Through our eyes, the universe is perceiving itself. Through our ears, the universe is listening to its harmonies. We are the witnesses through which the universe becomes conscious of its glory, of its magnificence."

*Alan Wilson Watts*

In this thesis we explore how eye tracking can be utilized to improve computer vision algorithms for images and videos and how advancement in higher level semantic understanding helps predict the eye tracking based salient regions in images. There

has been significant advancement in state-of-the-art eye tracking technology and recent eye trackers have become affordable, easy to use and more accurate. It has become practically feasible to collect eye tracking data when subjects are consuming multimedia content from digital displays such as computer monitors, tablets etc. without any constraints on head position, and availability of this additional data can greatly benefit computer vision and multimedia algorithms. Also, human attention is naturally biased towards high level semantic objects in images and videos, therefore, object extraction and annotation algorithms can utilize this additional information to improve the performance of state-of-the-art.

We explore four problems in this thesis relevant to this theme. In the third chapter, we explore how object detection in images can benefit from the availability of eye tracking data. We learn face and text eye tracking priors by only analyzing eye tracking data which reduces the search space for state-of-the-art faces and text detectors. This ensures significant gain in precision at negligible loss in recall. In addition, the proposed approach is the first effort to predict image categories from eye tracking data alone. We are able to predict face and text regions by only analysing eye tracking data from multiple subjects. In the following fourth chapter, we extend the object detection idea to object extraction in videos. Here, we utilize a more generic framework, not restricted to face and text categories. Typically, object segmentation algorithms in video focus on extracting moving objects, however motion might not be the only criterion

which captures the importance of an object in a scene. We believe visual attention is a better indicator of saliency of an object in a video sequence and the proposed approach aims to extract objects which capture a significant portion of visual attention in a scene. The proposed approach utilizes a two step process similar to the previous chapter, where we extract visual tracks representing visually salient objects in the scene which eventually guides an object search algorithm. The object search algorithm combines objectness measure in a novel multiple object extraction framework on a mixed graph optimized using integer programming. The proposed algorithm extracts more meaningful objects compared to algorithms without eye tracking prior and outperforms state-of-the-art which use eye tracking fixation prior.

In the fifth chapter we propose a saliency based algorithm which reduces the search space for text detectors. The algorithm learns a text attention map using a support vector machine, from multiple saliency algorithms especially focusing on regions where text detection algorithm fails. This saliency guided search approach improves over state-of-the-art, however as expected its performance gain is lower than what one would expect using eye tracking data as in chapter 3. In the sixth chapter we address the problem of predicting where people look, represented as visual attention map in images for light field images. We specifically investigate the role of two contextual factors, camera focus and object co-occurrence to predict the visual attention map. The state-of-the-art algorithms combined low, mid and high-level features using a linear support vector ma-

chine to predict the visual attention map. In addition to these features, we propose novel scene context features which combines object co-occurrence and camera focus based features and use a regression algorithm to predict the visual attention map. The proposed algorithm outperforms state-of-the-art saliency and visual attention algorithms. In addition, we also introduce a novel application to predict the best image from a set of Lytro light field images based on how visual attention overlaps with focus.

## 7.1 Future Work

As eye tracking enabled computer vision is in a fairly nascent stage, there are a plethora of open problems in this field. Most of the problems in this thesis require eye tracking data from multiple subjects, where practical applications such as viewing multimedia benefit from. Recently there has been significant advancement in wearable technology. In scenarios where eye trackers will be coupled with smart glasses, eye tracking data from only a single user is available. Therefore, augmenting the viewing area with useful information stemming from eye tracking data would have several practical applications. Therefore, developing algorithms which only require a single user's eye tracking data for scene understanding has immense potential applications. Additionally, related to visual attention modeling, in many cases obtaining a single person's eye tracking data is easier. Therefore, one can explore the feasibility of predicting the

visual attention map from a single user's attention map by adapting the single user's attention map using the image information.

Eye tracking is a convenient way to provide weak supervision and it would be interesting to investigate the utility of eye tracking data in other computer vision algorithms such as video based activity recognition and image retrieval. Various activity patterns might have signature eye movements and can help improve the performance of state-of-the-art. Also, eye tracking data can pick up critical aspects of the image which retrieval algorithms can primarily focus on.

We also suggest some improvements to the proposed techniques which can widen its applicability. The algorithm proposed in chapter 3 is designed to mitigate the false positives coupled with negligible loss in false negatives. In several computer vision applications such as face detection in surveillance feeds, where traditional detection algorithms will fail due to the quality of the video stream, eye tracking data can be used to improve recall by reducing the false negative rate. The design of such algorithms should also ensure minimal loss in precision. The proposed methods require the eye tracking data to be present both in the training and the test phases. It will be interesting to explore problems in a relatively restricted setting such as surveillance videos, where eye tracking data is available only in the training phase. The interaction between the eye tracking data and computer vision application can be learnt which can potentially improve the results in the test phase. Finally, it would be interesting to investigate the

utility of other high level semantic factors which affect visual attention and eventually bridge the gap between human attention and predicted visual attention maps.

# Bibliography


[1]

[2]

[3] Eyelink 1000. *http://www.sr-research.com/EL_1000.html/*. Accessed: 2013-19-03.

[4] Eyetribe eye tracker. *https://theeyetribe.com/*. Accessed: 2014-31-08.

[5] Lytro images. *https://pictures.lytro.com/*. Accessed: 2012-06-07.

[6] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2189–2202, 2012.

[7] Bogdan Alexe, Nicolas Heess, Yee Whye Teh, and Vittorio Ferrari. Searching for objects driven by context. In *Advances in Neural Information Processing Systems 25*, pages 890–898, 2012.

[8] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[9] A. Borji, L. Itti, J. Liu, P. Musialski, P. Wonka, J. Ye, S. Ji, W. Xu, M. Yang, K. Yu, et al. State-of-the-art in visual attention modeling.

[10] Ali Borji, Dicky N Sihite, and Laurent Itti. Salient object detection: A benchmark. In *Computer Vision–ECCV 2012*, pages 414–429. Springer, 2012.

[11] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(11):1222–1239, 2001.

[12] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[13] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 2009.

[14] Andreas Bulling and Hans Gellersen. Toward mobile eye-based human-computer interaction. *Pervasive Computing, IEEE*, 9(4):8–12, 2010.

[15] Andreas Bulling, Daniel Roggen, and Gerhard Tröster. *Wearable EOG goggles: eye-based interaction in everyday environments*. ACM, 2009.

[16] Andreas Bulling, Jamie A Ward, Hans Gellersen, and Gerhard Troster. Eye movement analysis for activity recognition using electrooculography. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):741–753, 2011.

[17] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12), 2009.

[18] Moran Cerf, Jonathan Harel, Alex Huth, Wolfgang Einhäuser, and Christof Koch. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. In *Attention in Cognitive Systems*, pages 15–26. Springer, 2009.

[19] C.C. Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[20] Kai-Yueh Chang, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. Fusing generic objectness and visual saliency for salient object detection. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 914–921. IEEE, 2011.

[21] Albert YC Chen and Jason J Corso. Propagating multi-class pixel labels throughout video frames. In *Image Processing Workshop (WNYIPW), 2010 Western New York*, pages 14–17. IEEE, 2010.

[22] Huizhong Chen, Sam S Tsai, Georg Schroth, David M Chen, Radek Grzeszczuk, and Bernd Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 2609–2612. IEEE, 2011.

[23] Xiangrong Chen and Alan L Yuille. Detecting and reading text in natural scenes. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–366. IEEE, 2004.

[24] J. Cohen and P. Cohen. *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum, 1975.

[25] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[26] Corinna Cortes and Vladimir Vapnik. Support vector machine. *Machine learning*, 20(3):273–297, 1995.

[27] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for multi-class object layout. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 229–236. IEEE, 2009.

[28] Nitin Dhavale and Laurent Itti. Saliency-based multifoveated mpeg compression. In *Signal processing and its applications, 2003. Proceedings. Seventh international symposium on*, volume 1, pages 229–232. IEEE, 2003.

[29] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1271–1278. IEEE, 2009.

[30] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18, 2008.

[31] Ian Endres and Derek Hoiem. Category independent object proposals. In *Computer Vision–ECCV 2010*, pages 575–588. Springer, 2010.

[32] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2963–2970. IEEE, 2010.

[33] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[34] Dmitry Fedorov, Baris Sumengen, and B. S. Manjunath. Multi-focus imaging using local focus estimation and mosaicking. In *IEEE International Conference on Image Processing 2006 (ICIP06)*, Oct 2006.

[35] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[36] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):989–1005, 2009.

[37] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. *Advances in neural information processing systems*, 17(481-488):1, 2004.

[38] Todor Georgiev, Zhan Yu, Andrew Lumsdaine, and Sergio Goma. Lytro camera technology: theory, algorithms, performance analysis. In *IS&T/SPIE Electronic Imaging*, pages 86671J–86671J. International Society for Optics and Photonics, 2013.

[39] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1915–1926, 2012.

[40] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph based video segmentation. *IEEE CVPR*, 2010.

[41] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Image Processing, IEEE Transactions on*, 19(1):185–198, 2010.

[42] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2014.

[43] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. 2007.

[44] Varsha Hedau, Derek Hoiem, and David Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *Computer Vision–ECCV 2010*, pages 224–237. Springer, 2010.

[45] C. Hickey and J. Theeuwes. Context and competition in the capture of visual attention. *Attention, Perception, & Psychophysics*, 73(7):2053–2064, 2011.

[46] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(1):194–201, 2012.

[47] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[48] Chang Huang, Bo Wu, and Ramakant Nevatia. Robust object tracking by hierarchical association of detection responses. In *Computer Vision–ECCV 2008*, pages 788–801. Springer, 2008.

[49] F. Imamoglu Konuskan. Visual saliency and biological inspired text detection. *Masters Thesis, Technical University Munich & California Institute of Technology*, 2008.

[50] Laurent Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318, 2004.

[51] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[52] C.D. Jones, A.B. Smith, and E.F. Roberts. Article title. In *Proceedings Title*, volume II, pages 803–806. IEEE, 2003.

[53] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[54] S Karthikeyan, Diana Delibaltov, Utkarsh Gaur, Mei Jiang, David Williams, and BS Manjunath. Unified probabilistic framework for simultaneous detection and tracking of multiple objects with application to bio-image sequences. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1349–1352. IEEE, 2012.

[55] S Karthikeyan, Vignesh Jagadeesh, and BS Manjunath. Learning bottom up text attention maps for text detection using stroke width transform. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 211–215. IEEE, 2013.

[56] S Karthikeyan, Vignesh Jagadeesh, and BS Manjunath. Learning top-down scene context for visual attention modeling in natural images. *ICIP, IEEE*, 2013.

[57] Wojtek J Krzanowski and WJ Krzanowski. *Principles of multivariate analysis*. Oxford University Press Oxford:, 2000.

[58] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[59] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. Key-segments for video object segmentation. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1995–2002. IEEE, 2011.

[60] Stan Z Li. *Markov random field modeling in computer vision*. Springer-Verlag New York, Inc., 1995.

[61] Simon M Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, and Robert Young. Icdar 2003 robust reading competitions. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, volume 2, pages 682–687, 2003.

[62] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 670–677. IEEE, 2012.

[63] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542. ACM, 2002.

[64] S.C. Mack and M.P. Eckstein. Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of Vision*, 11(9), 2011.

[65] Kazuki Maeno, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Light field distortion feature for transparent object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2786–2793. IEEE, 2013.

[66] Vijay Mahadevan and Nuno Vasconcelos. Saliency-based discriminant tracking. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1007–1013. IEEE, 2009.

[67] Stefan Mathe and Cristian Sminchisescu. Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012*, pages 842–856. Springer, 2012.

[68] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active segmentation with fixation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 468–475. IEEE, 2009.

[69] Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch

prior. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 22–28. IEEE, 2012.

[70] Frank Moosmann, Diane Larlus, and Frederic Jurie. Learning saliency maps for object categorization. 2006.

[71] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[72] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007.

[73] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *Computer Vision–ECCV 2014*, pages 361–376. Springer, 2014.

[74] Omkar M Parkhi, Andrea Vedaldi, CV Jawahar, and Andrew Zisserman. The truth about cats and dogs. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1427–1434. IEEE, 2011.

[75] Omkar M Parkhi, Andrea Vedaldi, A Zisserman, and CV Jawahar. Cats and dogs. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3498–3505. IEEE, 2012.

[76] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *Computer Vision–ECCV 2010*, pages 30–43. Springer, 2010.

[77] Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1454–1461. IEEE, 2009.

[78] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision research*, 39(19):3157–3163, 1999.

[79] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314. ACM, 2004.

[80] Ueli Rutishauser, Dirk Walther, Christof Koch, and Pietro Perona. Is bottom-up attention useful for object recognition? In *Computer Vision and Pattern*

*Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–37. IEEE, 2004.

[81] Karthikeyan S, Vignesh Jagadeesh, Renuka Shenoy, Miguel Eckstein, and B.S. Manjunath. From where and how to what we see. In *Computer Vision, 2013 IEEE International conference on*. IEEE, 2013.

[82] Palaiahnakote Shivakumara, Trung Quy Phan, and Chew Lim Tan. A laplacian approach to multi-oriented text detection in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(2):412–419, 2011.

[83] Christian Siagian and Laurent Itti. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on*, 25(4):861–873, 2009.

[84] E.P. Simoncelli and W.T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Image Processing, 1995. Proceedings., International Conference on*, volume 3, pages 444–447. IEEE, 1995.

[85] A.B. Smith, C.D. Jones, and E.F. Roberts. Article title. *Journal*, 62:291–294, January 1920.

[86] Ramanathan Subramanian, Victoria Yanulevskaya, and Nicu Sebe. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 33–42. ACM, 2011.

[87] Jin Sun and Haibin Ling. Scale and object aware image retargeting for thumbnail browsing. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1511–1518. IEEE, 2011.

[88] Q. Sun, Y. Lu, and S. Sun. A visual attention based approach to text extraction. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3991–3995. IEEE, 2010.

[89] Salil Tambe, Ashok Veeraraghavan, and Amit Agrawal. Towards motion-aware light field video for dynamic scenes. 2013.

[90] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

[91] David Tsai, Matthew Flagg, and James M.Rehg. Motion coherent tracking with multi-label mrf optimization. *BMVC*, 2010.

[92] Eleonora Vig, Michael Dorr, and David Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Computer Vision–ECCV 2012*, pages 84–97. Springer, 2012.

[93] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[94] Tina Walber, Ansgar Scherp, and Steffen Staab. Can you see it? two novel eye-tracking-based measures for assigning tags to image regions. In *Advances in Multimedia Modeling*, pages 36–46. Springer, 2013.

[95] Dirk Walther, Laurent Itti, Maximilian Riesenhuber, Tomaso Poggio, and Christof Koch. Attentional selection for object recognitiona gentle way. In *Biologically Motivated Computer Vision*, pages 472–479. Springer, 2002.

[96] Kai Wang and Serge Belongie. Word spotting in the wild. In *Computer Vision–ECCV 2010*, pages 591–604. Springer, 2010.

[97] Peng Wang, Jingdong Wang, Gang Zeng, Jie Feng, Hongbin Zha, and Shipeng Li. Salient object detection for searched web images via global saliency. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3194–3201. IEEE, 2012.

[98] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009.

[99] Sven Wanner and Bastian Goldluecke. Globally consistent depth labeling of 4d light fields. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 41–48. IEEE, 2012.

[100] Sven Wanner, Christoph Straehle, and Bastian Goldluecke. Globally consistent multi-label assignment on the ray space of 4d light fields. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1011–1018. IEEE, 2013.

[101] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[102] A. L. Yarbus. Eye movements and vision. *Plenum, New York*, 1967.

[103] Qixiang Ye, Qingming Huang, Wen Gao, and Debin Zhao. Fast and robust text detection in images and video frames. *Image and Vision Computing*, 23(6):565–576, 2005.

[104] Chucai Yi and YingLi Tian. Text string detection from natural scenes by structure-based partition and grouping. *Image Processing, IEEE Transactions on*, 20(9):2594–2605, 2011.

[105] Junyong You, Guizhong Liu, Li Sun, and Hongliang Li. A multiple visual models based perceptive analysis framework for multilevel video summarization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(3):273–285, 2007.

[106] Dong Zhang, Omar Javed, and Mubarak Shah. Video object segmentation through spatially accurate and temporally dense extraction of primary object regions. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 628–635. IEEE, 2013.

[107] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[108] Yu Zhong, Hongjiang Zhang, and Anil K. Jain. Automatic caption localization in compressed video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(4):385–392, 2000.