UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Foveated Vision Models for Search and Recognition

Permalink

https://escholarship.org/uc/item/6m78q2j7

Author

Ngo, Thuyen Van

Publication Date

2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Foveated Vision Models for Search and Recognition

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

Doctor of Philosophy
in
Electrical and Computer Engineering

by

Thuyen Van Ngo

Committee in charge:

Professor B.S. Manjunath, Chair
Professor Miguel Eckstein
Professor Michael Liebling
Professor Kenneth Rose

December 2018

The Dissertation of Thuyen Van Ngo is approved.

Professor Miguel Eckstein

Professor Michael Liebling

Professor Kenneth Rose

Professor B.S. Manjunath, Committee Chair

September 2018

Foveated Vision Models for Search and Recognition

Copyright © 2018

by

Thuyen Van Ngo

To my parents, my sister, and my wife for all the sacrifices you have made.

Acknowledgements

I would like to thank my advisor Professor Manjunath for the great opportunity to work at the Vision Research Lab at the department of Electrical and Computer Engineering. He has always been patient and has given me the great flexibility to explore many areas. His help and guidance are invaluable for my research at UCSB. Without his sponsorship it would not be possible for me to complete my graduate study. I am also indebted to Professor Eckstein for his patience in teaching me understand human vision and helping me with the human subjects experiments. I want to thank Professor Rose, Professor Liebling for serving as my committee members, providing critical suggestions and comments on my research.

Thanks to all the members of the Vision Research Lab for the companionship and support. I have learned a lot from Karthik on problem formulation and presentation. I have thoroughly enjoyed many research ideas through discussions with Archith, Utkarsh, Amir, Wenhui, Oytun and PoYo. Then, everyone else provided both friendship and advice: Carlos, Nilou, Chris Wheat, Tom and Lingyun. Special thanks to the perception team at Magic Leap, Inc. for offering me the summer internship and teaching me a lot on the practical side of computer vision.

Last, but not least, I cannot express how grateful I am to my wife Lien

for her emotional support and constant understanding throughout my years at UCSB.

Curriculum Vitæ

Thuyen Van Ngo

Education

2018	Doctor of Philosophy Department of Electrical and Computer Engineering University of California, Santa Barbara.
2013	Master of Science Department of Electrical and Computer Engineering University of California, Santa Barbara.
2010	Bachelor of Science Department of Electrical Engineering Hanoi University of Technology, Vietnam

Publications

How Do Drivers Allocate Their Potential Attention? Tao Deng, Hongmei Yan, Thuyen Ngo, B.S. Manjunath. IEEE Transactions on Neural Networks and Learning Systems, under review.

Brain Tumor Segmentation and Tractographic Feature Extraction from Structural MR Images for Overall Survival Prediction. PoYu Kao, Thuyen Ngo, Angela Zhang, B.S. Manjunath. MICCAI BraTS 2018, arXiv:1807.07716.

The unique face-centered human strategy for searching for people in the wild. Miguel Eckstein, Thuyen Ngo, and B.S. Manjunath. Vision Sciences Society Meeting 2018.

Optimizing Region Selection for Weakly Supervised Detection. Wenhui Jiang, Thuyen Ngo, B.S. Manjunath and Fei Su. arXiv:1708.01723

Saccade Gaze Prediction Using A Recurrent Neural Network. Thuyen Ngo, and B.S. Manjunath. IEEE International Conference on Image Processing 2017.

Eye tracking assisted extraction of attentionally important objects from videos. S. Karthikeyan, Thuyen Ngo, Miguel Eckstein and B.S. Manjunath. IEEE Conference on Computer Vision and Pattern Recognition 2015.

Experience

09/2011-08/2018	Research and Teaching Assistant Department of Electrical and Computer Engineering University of California at Santa Barbara.
06/2016-09/2016	Research Intern Computer Vision Team Magic Leap, Inc., Mountain View.

Abstract

Foveated Vision Models for Search and Recognition

by

Thuyen Van Ngo

Computer vision has made a significant progress in recent years thanks to advancement in neural network architectures and computing power. At the sensory level, the current machine vision systems sample the visual data uniformly to make predictions about the scene. This is in contrast with the human vision system that has high visual acuity only in a small central region, the fovea, and much coarser sampling away from the center. There has been a renewed interest, particularly in the context of active vision for robotics navigation and scene exploration, to develop biologically motivated methods that can leverage such foveated computations. While foveated vision offers computational savings at or near the region of interest, it requires eye movements to scan the scene for effective image understanding. The hypothesis is that methods that can leverage non-uniform sampling of the field of view together with eye-movements will lead to a new class of active vision systems that are optimized computationally for specific tasks of interest.

Inspired by the above observations, this research provides, for the first time, a comprehensive study of the human visual search in the constrained setting of person identification in the wild. A novel video database is created that systematically tests how different parts of a person contribute towards eye-movements and person identification. Our study shows that the search errors can dominate the overall recognition accuracy in human subject experiments. This calls for new strategies for integrating eye tracking with foveated image representations. Towards this two specific approaches are investigated further.

In the first approach, a deep neural network based method is developed to model eye movements. Using the long-short-term-memory to model the successive fixations. The proposed method outperforms state of the state of the art performance while simplifying the feature extraction procedure. The second approach focuses on the foveated image model that leverages multiple fixations. A convolutional neural network method is proposed that works directly with the foveated input images that achieves competitive recognition rates compared to standard neural networks operating on the same number of input pixels.

Overall the thesis investigates the requirements and implementations that could support active foveated vision, and lays down the ground work for future studies in this area.

Contents

Acknowledgements **v**

Curriculum Vitae **vii**

Abstract **ix**

1 Introduction **1**
1.1 Motivation 1
1.2 Summary of Dissertation Contributions 3

2 Visual Search and Active Vision **8**
2.1 Studies of Visual Attention and Search 8
2.2 Computational Models of Human Visual System 11
2.3 Foveated Active Vision 15

3 Face-centered human strategy for searching for people in the wild **17**
3.1 Introduction 18
3.2 Materials and Methods 20
3.3 Results 35
3.4 Discussion 47

4 Predicting fixation points with a recurrent neural network **51**
4.1 Introduction 52
4.2 Related Work 55
4.3 Approach 59
4.4 Experiments 67
4.5 Conclusion 73

5 Foveated Image Recognition **75**
5.1 Introduction 76
5.2 Related Work 77
5.3 Foveated Representation 78
5.4 Network Architecture 83
5.5 Experiments 91
5.6 Discussion 93

6 Conclusion and Future Work **95**
6.1 Conclusion 95
6.2 Future Work 98

Bibliography **101**

Chapter 1

Introduction

1.1 Motivation

Different from machine vision, human vision possesses a movable spatially variant visual sensory system that has high resolution at the central fovea and decreasing resolution in the periphery. This allows us to process fine details where necessary (in the fovea) but still gives us enough information in the periphery for further exploration. The configuration reduces the number of pixels in the retina one thousand times less than the representation which uses high resolution in the whole field view. The pixel reduction benefits both communication and computation, allowing us to understand the scene very efficiently. Given intensive computation required in current computer vision algorithms, it is desirable to take this into account for machine vision systems. However, most computer vision algorithms only work with uniformly sampled images. Not until recently, there exits machine vision systems that can work with nonuniform sampled inputs [1, 2] as well as active vision systems [3, 4] and mechanical systems to support sensory movement [5, 1].

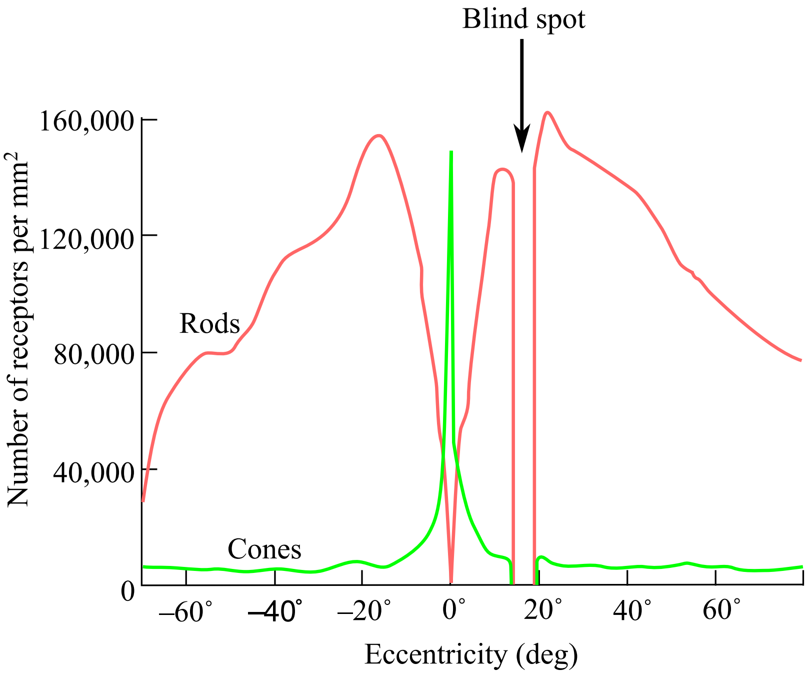


Figure 1.1: Distribution of rods and cones (photoreceptors) in the human retina, adapted from R. W. Rodieck, The First Steps of Seeing, Sinauer Associates, 1998 [6]. The graph illustrates that cones (green) have low density throughout the retina, with a sharp peak in the center of the fovea and rods (red) have high density throughout most of the retina, with a sharp decline in the fovea.

This trend is likely to continue with robotics systems requiring exploration and environment understanding at the same time. Given the similar goals of such machine vision systems and human vision, and the efficiency of human vision in carrying out common search and recognition tasks, it is desirable to understand how foveated vision work and build computational recognition system for foveated sensory signals directly.

1.2 Summary of Dissertation Contributions

Different from current computer vision systems where the input images are assumed to be collected in a regular 2-D grid, the foveated configuration requires the system to 1) plan sequence of locations to fixate; 2) gather information at each fixation; and 3) integrate acquired information across those locations. Our goal is to understand characteristics of such systems and build modern neural architectures that operate on the learned principles. Towards this, we first study how humans perform a visual search task. Next, a neural network based approach is proposed to model eye movements that further improves upon the current state of the art models. Finally, a new foveated convolutional neural network is presented that leverages multiple fixations and is adaptive to available computational resources. The main contributions are now summarized:

- This research provides new insights into visual search by humans. A novel and unique dataset is curated that includes videos in the wild. One interesting observation from the human subject experiments is that humans can make more search errors than recognition errors in person identification tasks. Also humans tend to look at below the face when the person's head is occluded or otherwise not visible, and look at the face regions even when the facial features are unavailable. Finally, current computational machine vision methods do not prioritize face regions as humans do in such person identification tasks.

- We propose a simple and efficient method for gaze estimation. The proposed approach removes dataset-dependent feature engineering steps and achieves state of the art performance.

- Complementing the gaze prediction, we propose a convolutional neural network for foveated image recognition. This is the first recognition system to operate on fovated images directly. The proposed method is able to handle multiple fixations, and can be made adaptive to the computational budget by allowing more fixations to improve the overall recognition performance.

This thesis is organized as follows:

- Chapter 2 discusses current research in visual search and foveated vision.

- Chapter 3 studies how humans search for target people in dynamic scenes. A novel video database is created that systematically tests how different parts of a person contribute towards eye-movements and person identification. In this task subjects rely strongly on the face where performance drops by a large amount when facial information is removed. Our study shows that the search errors can dominate the overall recognition accuracy in human subject experiments. The recorded eye movements show that humans have strong a bias towards faces. But when being forced to fixate at faces subjects do not obtain maximum performance, which suggests that face-centered strategy does not necessarily maximize the person identification performance in human subject trials but likely arises as a byproduct of the implementation of a heuristic strategy that optimizes perceptual performance across a battery of evolutionary important tasks. Performance of two current computer models, a foveated ideal observer and a naive convolutional neural network, is compared against human, showing that machine models treat faces similar to other features and are outperformed by human subjects by a large margin.

- Motivated by the results in Chapter 3, Chapter 4 aims to model human eye movement directly. Given an image, we want to predict the most likely sequence of fixations a human would follow. We leverage recent advances in image recognition using convolutional neural networks and sequence modeling with recurrent neural networks. Feature maps from convolutional neural networks are used as inputs to a recurrent neural network. The recurrent neural network acts like a visual working memory that integrates the scene information and outputs a sequence of fixations. The model is trained on human eye tracking data. The proposed approach removes dataset-dependent feature engineering steps and achieves state of the art performance.

- Chapter 5 goes a step further to build a model of image recognition that can operate on foveated input images. Current methods use spatially variant filtering to create foveated images, retaining the same number of pixels as the original inputs. Assuming a log-polar representation of foveated sensory signals, a circular convolutional neural network is designed to perform image recognition. To the best of our knowledge, this is the first time that a convolutional network is developed to work directly with the foveated image data. The proposed method is also able to handle multiple fixations giving better performance with more fixations, thus adaptive to the given computational budget.

- Chapter 6 summarizes the thesis and proposes future directions.

Chapter 2

Visual Search and Active Vision

Biological vision has always been a great source of inspiration for design of computer vision algorithms. Previous research includes methods that functionally mimic biological vision systems to varying degrees, to models that are primarily developed to explain biological observations. In the following we briefly review some well-known models of visual attention and search.

2.1 Studies of Visual Attention and Search

There exist many works on visual attention that estimate saliency or the gaze distribution over an image. Most methods rely on bottom-up processing, which is the processing of information that uses only the input from the environment. In [7] center-surround feature maps are computed from the Gaussian pyramid for color, intensity and orientation channels, and then combined into a single saliency map. Graph-Based Visual Saliency (GBVS) model [8] extracts similar features as [7] but builds a graph associated with each feature map. Saliency maps are then the stationary distributions of a Markov chain induced by the graphs, and finally combined into a single saliency map. The model proposed in [9] is based on center-surround property, contrast sensitivity function, visual masking and perceptual decomposition. Figure 2.1 shows the result of such a model beside the original image. The input image on the left contains two baskets of objects. The resulting saliency on the right is able to pick up regions with dominant objects in the scene.

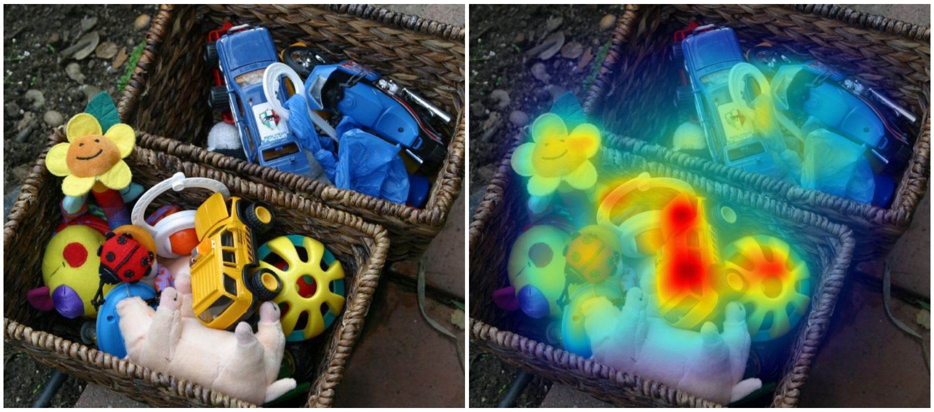
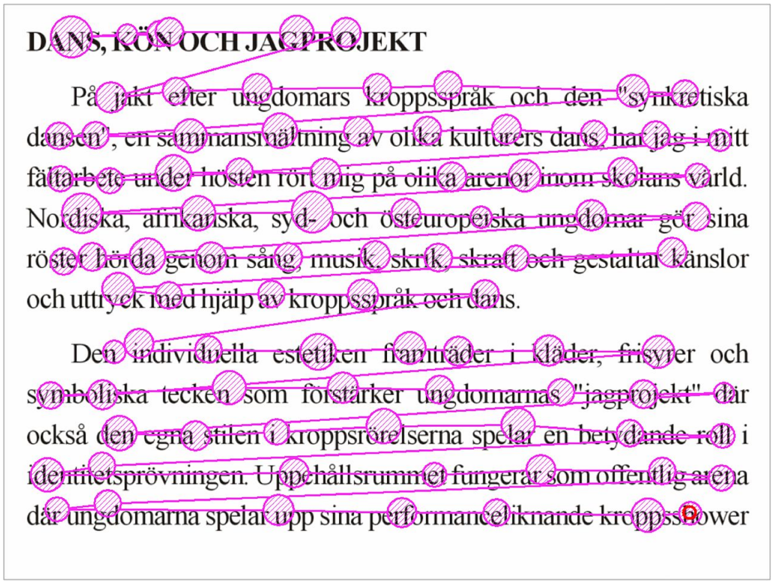


Figure 2.1: Example results of a visual attention model, obtained using algorithm from [9]. The input image on the left contains two baskets of objects. The heat map of output saliency is overlaid on the input image. The resulting saliency on the right is able to pick up regions with dominant objects in the scene.

Another body of work has devoted to visual search. Since search comes with goals, humans utilize top-down processing, taking advantage of prior knowledge about the targets to efficiently identify them. An importance question during search is the strategies humans use to make eye movements [10, 11, 12, 13]. Due to the foveated nature of the visual system, humans need to fixate to the object of interest to obtain a high resolution image of the stimuli in the fovea. Thus eye movements are needed to gaze at different objects in the scene. Such movements are called saccadic eye movements. Not only search but many other tasks would require top-down processing and making eye movements. Figure 2.2 shows typical saccades while reading a piece of text. The size of the circles represents the time spent at any one location and the line connecting any two circles represents a saccade.



DANS, KÖN OCH JAGPROJEKT

På jakt efter ungdomars kroppsspråk och den "synkretiska dansen", en sammansmältning av olika kulturers dans, har jag i mitt fältarbete under hösten rört mig på olika arenor inom skolans värld. Nordiska, afrikanska, syd- och östeuropeiska ungdomar gör sina röster hörda genom sång, musik, skrik, skratt och gestaltar känslor och uttryck med hjälp av kroppsspråk och dans.

Den individuella estetiken framträder i kläder, frisyrer och symboliska tecken som förstärker ungdomarnas "jagprojekt" där också den egna stilen i kroppsrörelserna spelar en betydande roll i identitetsprövningen. Uppehållsrummet fungerar som offentlig arena där ungdomarna spelar upp sina performanceliknande kroppsshower

Figure 2.2: Example of eye movement during reading obtained from [14].

2.2 Computational Models of Human Visual System

Most models of the human visual system focus on visual recognition, which is the ability to label an image with a meaningful category irrespective of viewing distance, position, size or context. For human vision, this task involves most of visual areas, from primary visual cortex (V1) to inferior temporal cortex (IT). Neurons in higher level can maintain selectivity while being invariant to positions and scales. IT neurons, for example, response strongly to various faces at different position and sizes but not at all to other stimuli. Current computational models achieved this ability by organizing computation in a hierarchical manner, corresponding each stage of computation with a neural process in cortex area. Computation at a certain stage pools or integrates different adjacent inputs from the previous stage, obtaining more complex representations. A computational unit, therefore, will be able to response similarly to slightly translated inputs. This invariance will increase with more stages of computation. A mong very fist hierarchical models are feedforward models with a homogeneous multilayered architecture. Later Fukushima proposed Neocognitron [15] architecture to further account for translation invariance. These models are all motivated from pioneering physiological studies by Hubel and Wiesel [16]. Many models have been proposed since then based on the simple-complex-cell models by Hubel and Wiesel, including VisNet [17], HMAX [18], and, convolutional neural networks (CNNs). A schematic of the HMAX model is shown in Figure 2.3.

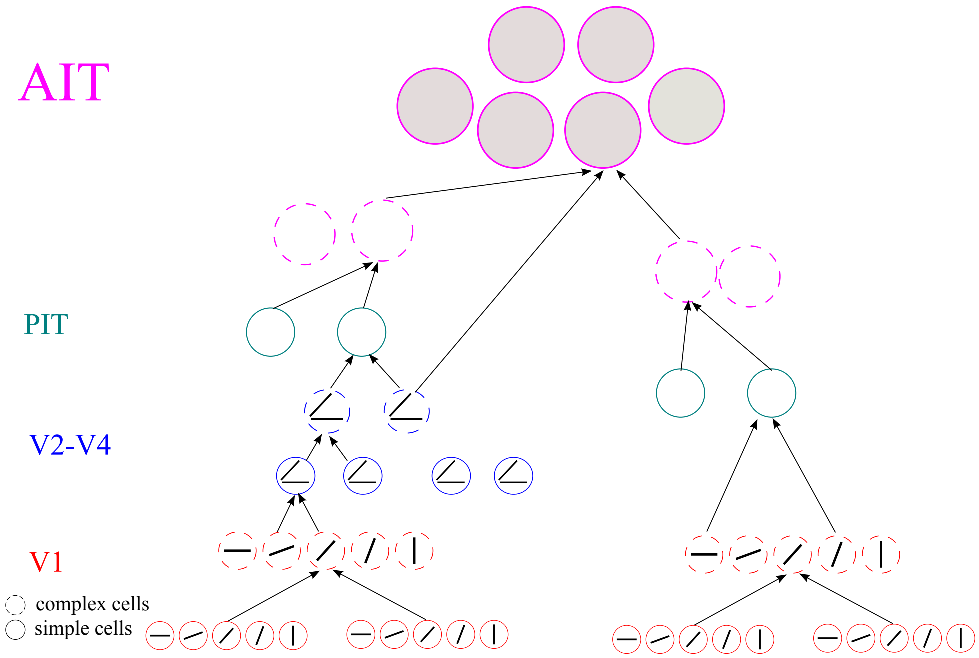


Figure 2.3: Schematic of the HMAX hierarchical computational model of the visual cortex [18]. V1, V2 and V4 correspond to primary, secondary and quaternary visual areas; PIT and AIT to posterior and anterior inferior temporal areas, respectively.

Variants of Fukushimas Neocognitron [15] have become popular in computer vision and are popularly referred to as convolutional neural networks (CNNs). While existing computer alogirthm algorithms could recognize geometric patterns in images, they were not able to generalize very well, or learn how those patterns might occur in other parts of the image. Fukushimas contributions take advantage of the shift invariance property of visual input for

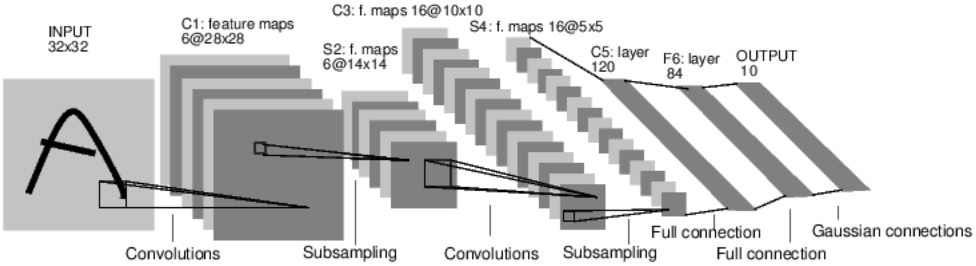


Figure 2.4: A modern convolutional architecture first proposed by Yann Lecun in [19] for digit recognition. The network consists of successive convolutions, nonlinearity and downsampling. At the end, activations are flattened into a vector before classification.

recognition build that into the classifiers. Yann Lecun then improved the architecture [19] and co-developed backpropagation for efficient training of such networks [20].

Another important feature of this architecture is that it forces weight-sharing among neurons and thus reduces the number of parameters involved. While traditional image classification approaches rely on manual encoding of features, having an expert define where certain patterns will occur in an image. CNNs can obtain this same information efficiently through training. Equipped with development of large datasets and unparalleled computing resources, CNNs have achieved state of the art performance in computer vision in recent years.

2.3 Foveated Active Vision

Active vision has been studied extensively in robotics for the ability to actively move the sensors to explore the environment. This ability is desirable for mobile robots where they need to navigate and understand unseen environments. Such a system shares many goals similar to human observers exploring a new environment. Motivated by efficiency of human visual system, many recent works have built hardware to actuate human-like foveated vision systems, mostly focus on the device's ability to 1) capture foveated images and 2) make cameras movable. Examples of such systems include foveated wide angle lens for active vision [21], binocular, foveated active vision system [4], reconfigurable foveated active vision system [22]. However, there are very limited image processing and computer vision methods that can operate on foveated inputs. Simple algorithms exist to detect lines, circles [23] and other simple shapes but they are very far away from being able to recognize object classes or guide camera movements for navigation.

In order to fully take advantage of the aforementioned foveated sensors, it is necessary to develop algorithms that can work directly with foveated inputs, including eye movement planing and object recognition. Inspired by the human visual system, this thesis will explore what would be important to humans and how they make eye movements 2) models of human eye movements

3) recognition models with foveated input images.

Chapter 3

Face-centered human strategy for searching for people in the wild

This chapter focuses on understanding how humans search for person identification in dynamic scenes. More specifically we study factors that would affect the search task, including where human subjects fixate in the scene while carrying out the task. In static scenes, current research suggests that human subjects focus on the face region when the resolution enables face recognition, and use the whole body information otherwise [24], [25], [26] and [27]. In contrast, the results presented in this chapter focus on dynamic, *in the wild*, data where there could be multiple distractors such as other people in the scene moving around in a natural, unconstrained, environment. Humans tend to look at the face region in natural scenes and also in manipulated scenes where faces are removed. We conclude the chapter evaluating two different machine models for the task: a foveated ideal observer (FIO) and a naive convolutional neural network (CNN). Both CNN and FIO do not weight faces as important as humans do.

3.1 Introduction

There is a large body of work investigating human visual search with simple synthetic displays (Ts among Ls) [28, 29, 30, 31, 32, 33]; or simple target in noise [34, 35, 36, 37, 38] or objects in real scenes [39, 40, 41] and trying to understand the underlying eye movement strategies [10, 11, 12, 13], important features [28, 42] and processing limitations. There is also a separate literature identifying the eye movement plans [43, 44, 45, 46, 47] and facial features utilized by humans to determine the identity [48], gender an emotion in a face. Those studies suggest that humans primarily use the eyes and mouth for the face judgments [48, 49] and fixate at a featureless point just below the eyes to optimize the acquisition of information through their foveated visual system [43]. A number of recent studies have expanded the processes by which a person is recognized from a picture or video of the whole person [24, 25, 27, 26]. Yet, those studies involve a single person in a picture or video rather than in a search in a crowd scenario and have not looked at the eye movement patterns of observers and their functional importance.

Little is known about how humans search for another person in crowds. Which features of a person (face, head, body) guide the eye movements towards searched people in the crowd? Which features are critical to correctly determine the identity once a target person is fixated? Do humans direct their eyes to a consistent location within a person and do these fixations have a functional importance for identification? And finally, does the human utilization of features and eye movements reflect an interaction between the distribution of visual information critical for the person identification and the foveated nature of the human visual system?

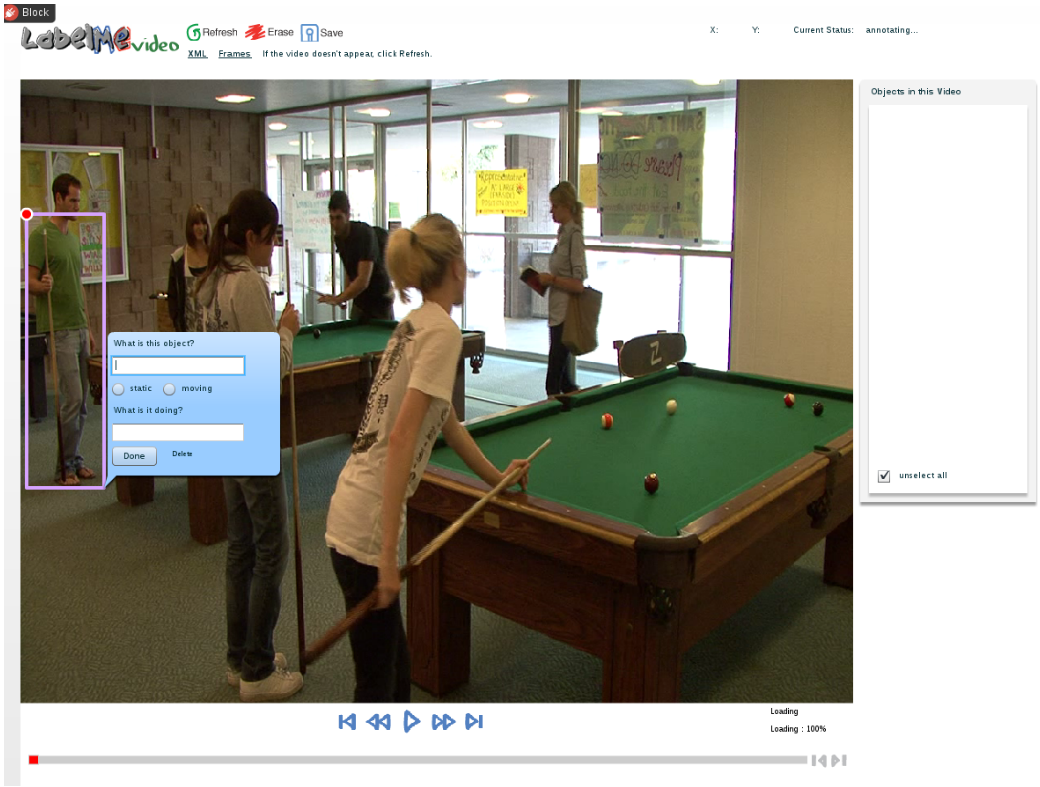
To answer these questions, we construct a novel video-in-the-wild database for human subject experiments. The subjects are tasked with deciding if a given video segment contains one of two target persons or neither, a 3-category decision. The database is curated so as to expose specific parts of the human body, i.e., face only, body only, and facial hair with body. We investigate the influence of the presence or absence of these features on the search error where the subjects fail to fixate on the target of interest, and separate that from the recognition errors which are due to mis-identification while fixating on the correct target. We measure the location of fixations within the head-to-toe image of people in the videos to assess whether there is a consistent point of fixation and how fixation varies with viewing angle as well as absence of salient features. To assess whether the position of fixation to people has a functional importance for perceptual performance, we conducted a separate psychophysical study with still pictures where observers are instructed to maintain fixation at different points along the persons body.

3.2 Materials and Methods

Subjects. Each separate study (free eye movements and forced fixation study) was completed by a separate group of 60 undergraduate students (120 total, 55 males and 65 females) participating for research credits. Informed consent was obtained for all subjects following guidelines provided by the Institutional Review Board at the University of California, Santa Barbara. Participants also provided consent to the utilization of their pictures in scientific publications.

Eye Tracking. The left eye of each participant was tracked using an SR-Research Eyelink 1000 Tower Mount eye tracker sampling at 1000 Hz. A nine-point calibration and validation was run before each 120-trial session, with a mean error of no more than 0.5 degrees of visual angle. If participants moved their eyes more than one degree from the center of the fixation cross before the stimulus was displayed, the trial would be aborted and restarted with a new stimulus.

3.2.1 Visual Search Task



If the video doesn't appear, click Refresh.

Current Status: annotating...

Loading

Loading : 100%

Figure 3.1: A snapshot of LabelMeVideo toolbox for video annotation. The software provides rectangle and polygon drawing tools for each frame. The annotator only needs to draw a shape (rectangle in this case) in sampled frames and the software will provide smoothed shapes across all frames by interpolation. The tool is used to annotate head and body.

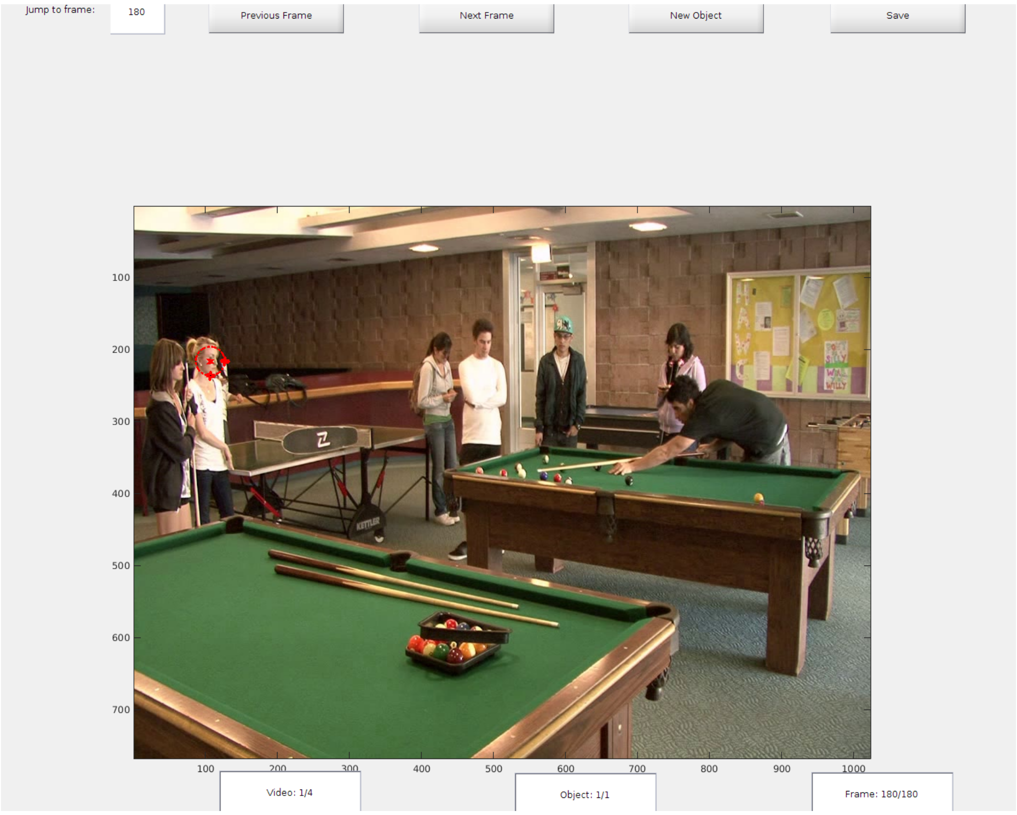


Figure 3.2: A snapshot of the custom face annotation tool in Matlab. This provides an easy way to draw ellipses to fit the shape of faces. In the above image, the ellipse is initialized on a person's face (red region on the left) and resized by the annotator accordingly to fit the face.

Stimuli. The dataset consists of 120 videos collected at UCSB campus by the Vision and Image Understanding Lab. The videos were staged with our target people and other distractor people at different campus locations across different days to ensure that clothing varied across the videos. The filming lasted over six weeks. All participants in the video have given written consent to be filmed and allow utilization of their images. Two people (a male and female) which we will refer to as Fando (male) and Lis (female) were assigned as targets. One third of the video clips contained Fando (but not Lis), one third contained Lis (but not Fando), and one third contained neither of them. Six second clips were extracted from each video for our experiments. To isolate certain features, individual frames need to be annotated with the corresponding features. Body and head annotations were conducted using LabelMeVideo [50]. The face annotation was done using an in-house software written in Matlab. All annotations were collected by undergraduate students at the University of California, Santa Barbara participating for research credits over a school year. The outline of the procedure is shown in Figure 3.1 and Figure 3.2. A background of each video can be obtained by taking median of all frames in the video due to the fixed camera settings. The region within each annotation is filled with this background erasing the specific feature while preserving immediate background. Sampled frames from stimuli can be viewed in Figure 3.3 and Figure 3.4.

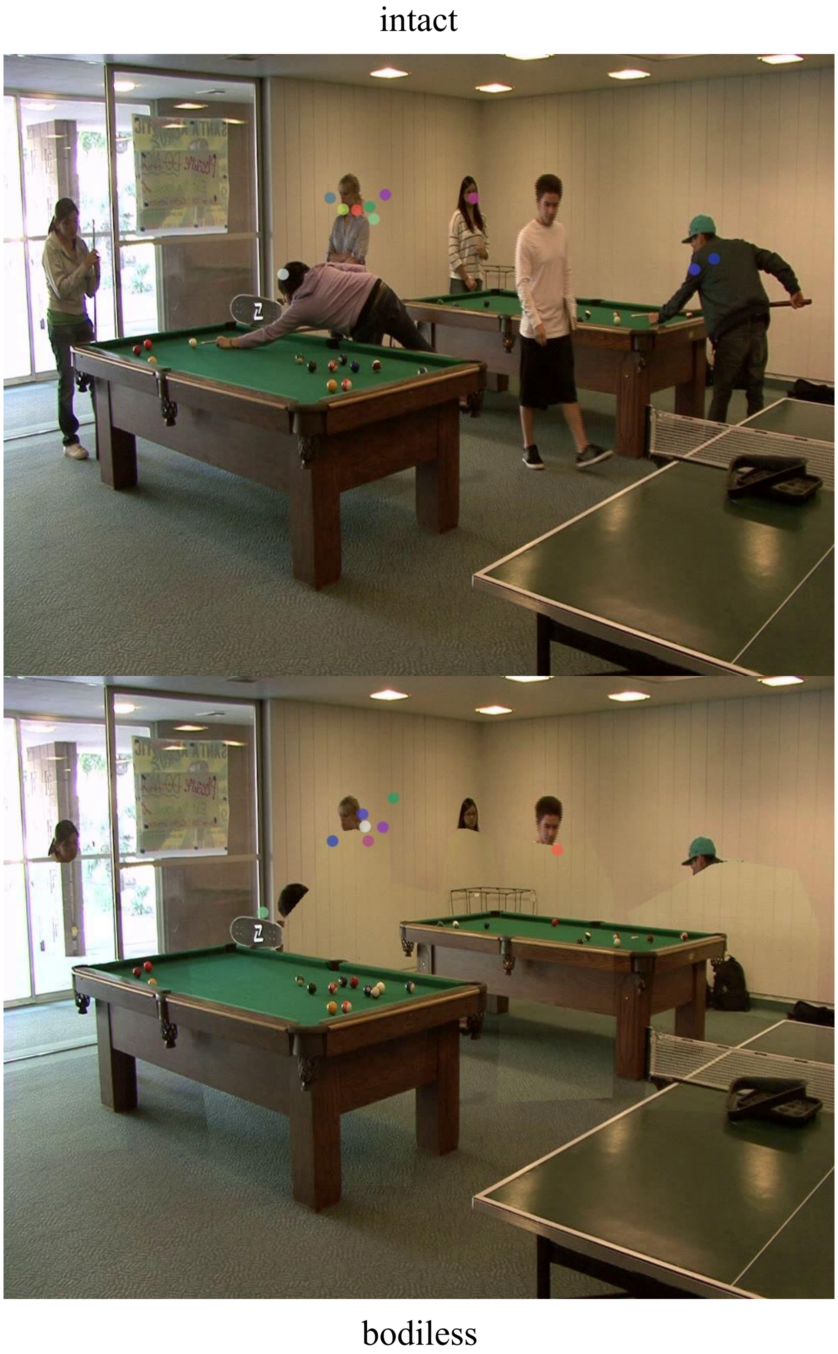
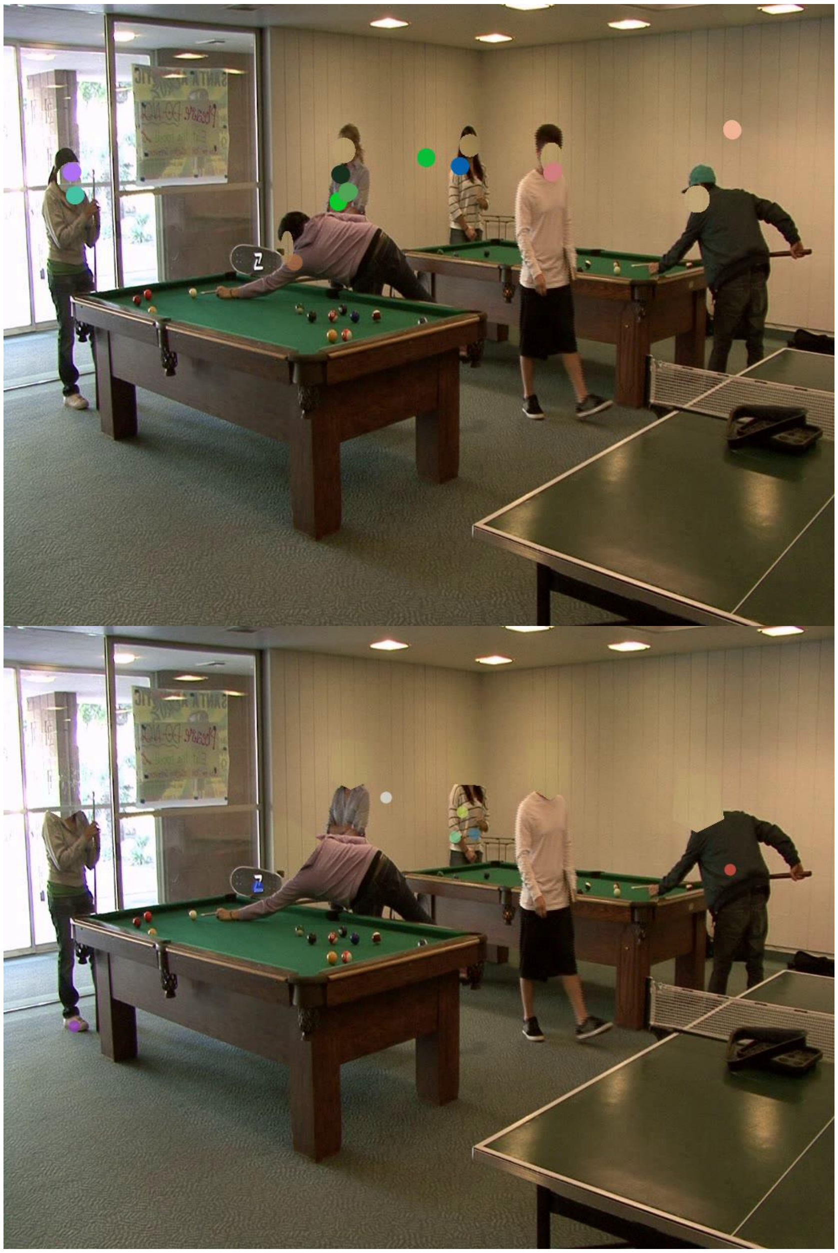


Figure 3.3: Example frames from four conditions in the experiment: intact (top) and bodiless (bottom). Sampled eye movement data from human subjects are also overlaid in the corresponding conditions.



faceless

headless

Figure 3.4: Example frames from four conditions in the experiment: faceless (top) and headless (bottom). Sampled eye movement data from human subjects are also overlaid in the corresponding conditions.

Task: The task was to assess whether Fando, Lis was present or neither were present (3 category task). There were four feature conditions intermixed: intact videos with all features (intact), headless, bodiless, and faceless. To minimize effects of memory of specific videos, the experiment was a between subject design, meaning that each observer saw a video only once and different videos might come from different conditions.

Procedure: Observers first watched sampled videos and images of targets to familiarize themselves with the targets (Fando and Lis). These sample images were not used in the experiment. They then were presented with a total of 120 video clips. During each trial, observers were briefly shown a video (framerate = 30 frames per second) with randomly chosen presentation times (1, 2, 3, 4, 5, or 6 sec). No specific instructions were given to observers about eye movements or search strategies. After the presentation of the video clip and response image was shown, observers select one of three keys to indicate whether Fando, Lis or neither were present. A schematic overview of the experiment is shown in Figure 3.5.

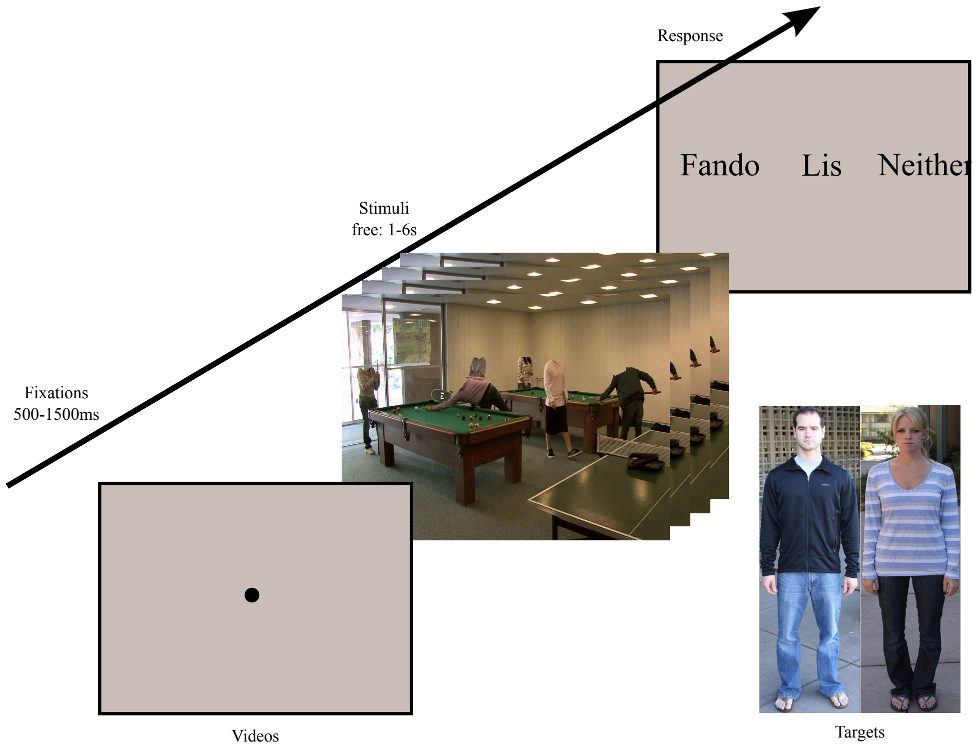


Figure 3.5: Task time line for the free eye movement experiment. First the participating human subject is asked to fixate to a specific fixation on the screen. A stimulus (a video) is then shown for a duration from one to six seconds. After that the subject needs to provide a response on where she/he sees the first target, the second target, or neither of them in the video.

3.2.2 Forced Fixation Task

Stimuli. 120 static images are cropped around a person of interest from the videos used in the visual search task. They are then normalized to a fixed scale occupying 15 degrees of visual angle. Similar to the visual search task, one third of the images contained Fando (but not Lis), one third contained Lis (but not Fando), and one third contained neither of them.

Task: The task was to assess whether either of Fando or Lis is present, or neither is present. There are four feature conditions intermixed: intact videos will all features (intact), headless, bodiless, and faceless. The experiment also uses a between subject design where each observer saw an image only once and different images might come from different conditions.

Procedure: Observers first watched sampled images of targets to familiarize themselves with the targets (Fando and Lis). These sample images were not used in the experiment. A dot representing the desired fixation location is then presented and the subject is asked to fixate to the dot. Stimulus is then presented for 200ms. The trial is discarded if the subject makes an eye movement away from the dot (1 degree). After the presentation of the image and response image was shown, observers select one of three boxes to indicate whether Fando, Lis or neither were present. A schematic overview of the experiment is shown in Figure 3.6.

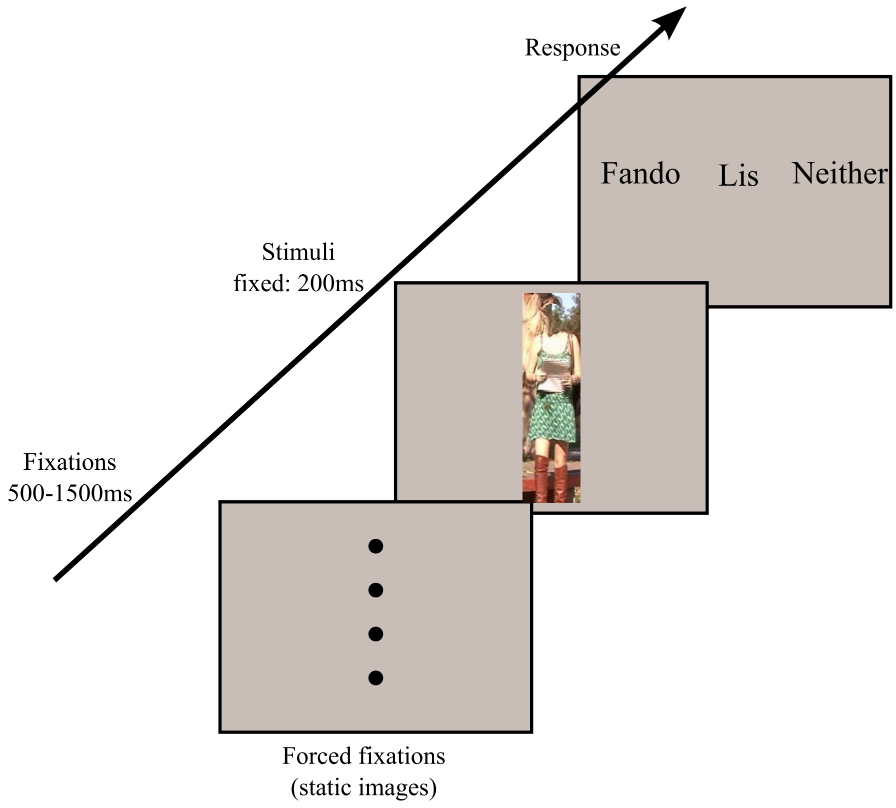


Figure 3.6: Task time line for the forced fixation experiment. First, the participating human subject is asked to fixate to one of four designated locations on the screen. The similus (an image) is shown for 200 ms. During the process, the subject is not allowed to make eye movements. If an eye movement is detected, the trial will be discarded. At the end the subject needs to provide a response on where she/he sees the first target, the second target or neither of them in the image.

3.2.3 Computational Models

We utilized computational models to generate theoretical predictions of the influence of features on recognition performance taking into consideration the distribution of discriminatory information across the head and body of people and the foveated nature of the human visual system. The models were developed for the forced fixation task with the still images extracted from the videos and utilized for the human study. Below we describe the two computational models.

Foveated Ideal Observer. The first model we utilized to evaluate what might be the optimal point of fixation for person identification is the foveated ideal observer. The model has been utilized previously to correctly predict the human optimal point of fixation to faces and how these change with central vision loss [43, 46, 51]. The model takes into account the varying spatial detail of visual processing with retinal eccentricity, and integrates the information across the visual field to make optimal decisions. To simulate the effects of eccentricity on sensitivity to different spatial frequencies, we used a spatially variant contrast sensitivity function (SVCSF) linear filtering function that took points of fixation, eccentricity, and direction away from fixation as variables.

The function has a form given by:

$$F(f, r, \theta) = c_0 f^{a_0} e^{-b_0 f - d_0(\theta) r^{n_0}} \tag{3.1}$$

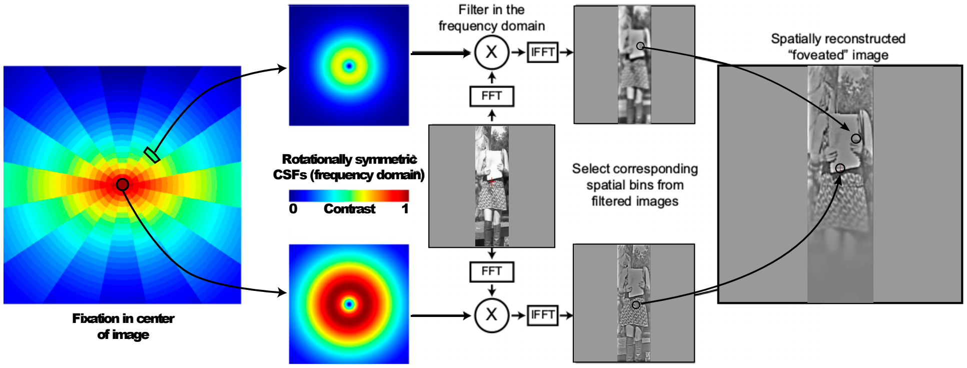
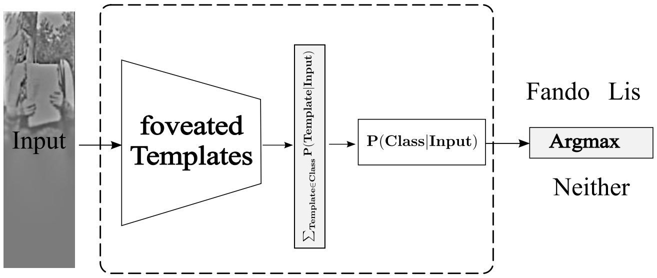
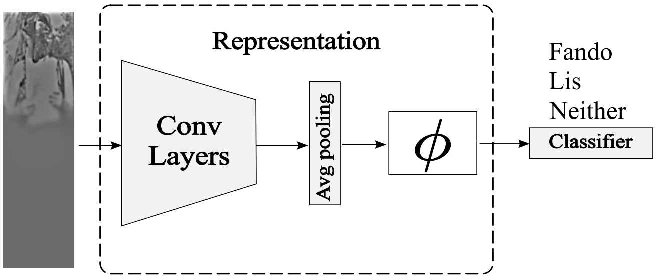

Filter in the frequency domain
X
IFFT
FFT
Spatially reconstructed “foveated” image
Rotationally symmetric CSFs (frequency domain)
0
Contrast
1
Select corresponding spatial bins from filtered images
FFT
X
IFFT
Fixation in center of image

Figure 3.7: Foveated (spatially varying) filtering. For a given fixation, the space is quantized into polar bins. For each bin, a filter with frequency response computed using Equation 3.1 is applied to the image within the corresponding bin. The results for all bins are combined in the output. The image demonstrate the procedure for two bins: one at the fixation and another bin nearby.

In the above equation f is spatial frequency, (r, θ) are polar coordinates of the considered location centered at the fixation. $d(\theta)$ is the orientation-dependent that is linearly interpolated among upward, downward and horizontal directions dd, du, dh. The parameters c_0, a_0, dd, du, dh of the equation are constants whose values are based on previous work for faces [43, 46, 51].


Input
foveated Templates
$\sum_{\text{Template}\in\text{Class}} \text{P}(\text{Template}|\text{Input})$
P(Class|Input)
Fando Lis
Argmax
Neither

(a) Overview of FIO prediction pipeline. A template is represented by a vector where each element is the dot product of the noisy template with other templates in the dataset. First a Gaussian distribution is built for each template assuming the noise follows a gaussian distribution. Given a noisy image, its likelihood of belonging to each class is computed using the aforementioned distribution.



(b) Overview of CNN prediction pipeline. First features are extracted using a pretrained CNN. An SVM classifier is then used to make prediction on target classes.

Figure 3.8: Foveated ideal observer and traditional convolutional neural network models for classification.

The exponential decay n_0 was fitted to match human experiment data in the forced fixation task. The overall filtering procedure is shown in Figure 3.8a. For any given fixation point, the input image (with the same contrast and additive white noise as viewed by the humans) is filtered by the SVCSF. The FIO compares this filtered noisy input with similarly filtered noise-free templates of each possible face, resulting in a set of template responses. The template responses follow a multivariate normal distribution with mean vector and covariance. The model then computes the multivariate normal likelihoods of all template responses given that each image of a target is present (Fando, Lis, or other people). The likelihoods are summed within each class, resulting in a collection of summed likelihood terms. The FIO then takes the maximum of these summed likelihoods as the decision. FIO model can be summarized in Figure 3.7. Since the same data is used for creating the model and measuring performance, FIO measure the goodness of fit rather than recognition performance.

We ran the FIO on a total of 120 still images are extracted from different videos. Of these images one third contained Fando, one third Lis and one third neither. White noise with a standard deviation of 4 was added to the images.

Convolutional Neural Networks. In contrast to FIO models, convolutional neural networks offer a way to validate the model for new images in the testing data set. Convolutional layers apply convolution operations to the input, passing the result to the next layer. The convolution emulates the response of an individual neuron to visual stimuli. The final layer acts as a classifier to produce the desired output. CNN model is summarized in Figure 3.8b.

We used Resnet152 [52] to extract features from images and only trained a Linear SVM [53] on the extracted features. Inputs to the CNN could be either original or foveated filtered images. Images are splits into 5-fold validation sets so that appearance of people are different among 5 sets. In each training session one fold is held out for validation and the results are averaged across 5 folds.

3.3 Results

3.3.1 Influence of features on human search accuracy

Figure 3.9 shows average perceptual performance (proportion correct, PC) of human subjects, for the feature conditions were 0.7892 ± 0.0118, 0.5717 ± 0.0144, 0.5492 ± 0.0144, 0.7300 ± 0.0128 for the intact, faceless headless and bodiless condition correspondingly. The results show a big influence on task performance for the face and head (24 % and 27 % PC reduction against the intact condition) compared to the body (5 % PC reduction), even though they are all statistically significant (p-value $\ll 0.05$). The contributions of head and face are less distinctive (p-value $= 0.055$). In detail, the pairwise two-sided t-test between conditions are: intact vs faceless $t = 12.8$, p-value $= 0$; intact vs headless $t = 14.50$, p-value $= 0$; intact vs bodiless $t = 3.4$, p-value $= 3.4 \times 10^{-4}$; bodiless vs headless $t = 10.9$, p-value $= 0$; bodiless vs faceless $t = 9.2$, p-value $= 0$; headless vs faceless $t = 1.6$, p-value $= 0.055$.

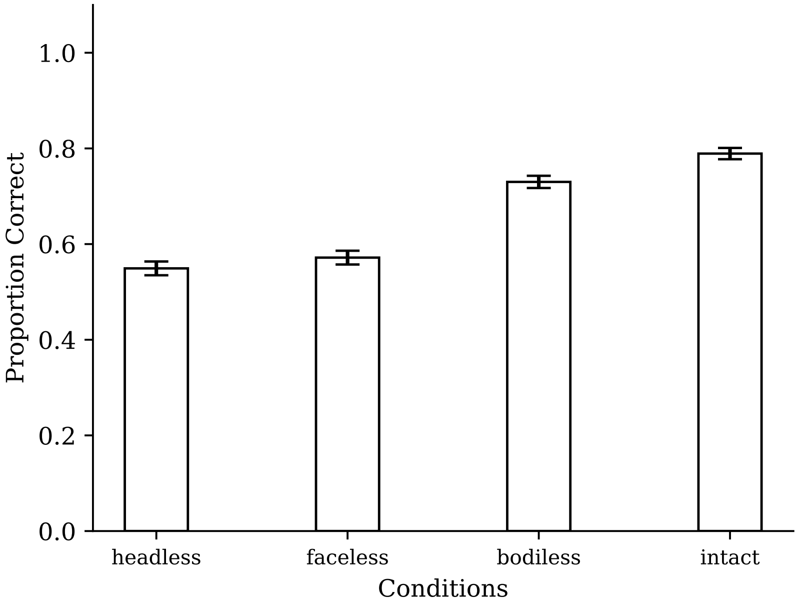


Figure 3.9: Human performance in Proportion Correct. Overall the face and head information are important for the search task, indicated by a large drop in human performance when face or head information are absent.

Figure 3.10 shows performance as a function of presentation time for the various feature conditions. Overall the pattern of results were similar across all presentation times with intact achieving the highest accuracy, bodiless, and then faceless and headless and no significant interaction (analysis of variance, or ANOVA); Grouping trials with the same viewing time we can see the average PC of all conditions are increasing overtime and their distinctions are less pronounced when videos are very short (at one second two-sided t-test between intact and bodiless p-value $= 0.33$ and faceless and headless p-value $= 0.7$) and become more obvious when time increases (at six seconds the test results are p-value $= 0.0034$ p-value $= 0.02$ correspondingly).

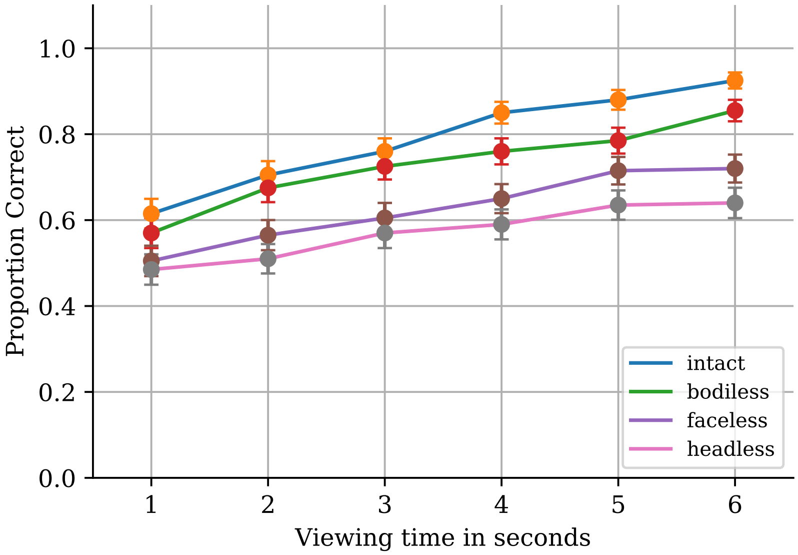


Figure 3.10: Human performance as a function of viewing time. The trends are similar to the overall performance 3.9, emphasizing the importance of face and head information.

3.3.2 Eye movement Analysis

Fixations during search. On average, each video contained an average of 3.7±0.14. Of all fixations, 45 % were directed towards people in the videos (40 % with a tolerance of 0.5 deg). Figure 3.11 shows the number of fixated people increased with presentation time but it was typically less than the total number of people in the video. Specifically the average numbers of fixated people are 1.7 ± 0.04, 2.645 ± 0.034, 2.795 ± 0.04, 2.99 ± 0.03, 3.29 ± 0.04 and 3.31 ± 0.04

for the viewing time of one, two, three, four, and five seconds correspondingly.

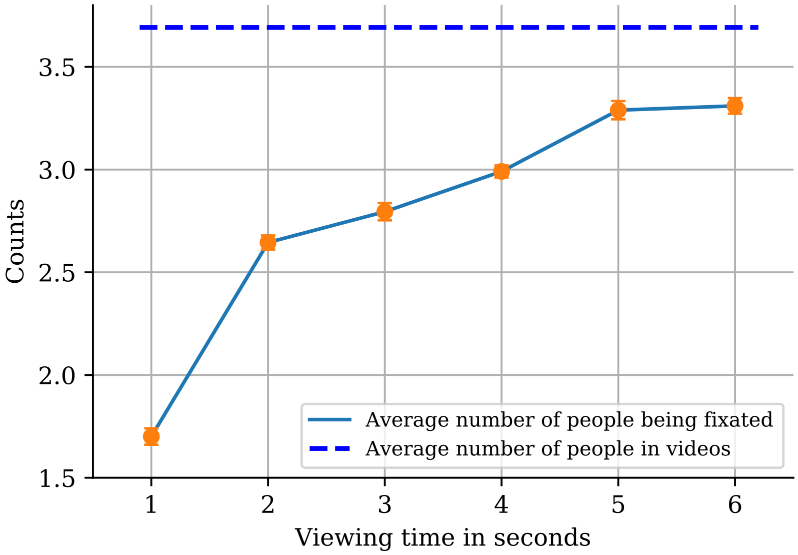


Figure 3.11: Number of people being fixated as a function of viewing time.

3.3.3 Search Error and Recognition Error

To isolate contributions of the features to eye movement guidance towards the targets from contributions to recognition performance, we divided errors into two categories. Search errors refer to trials in which observers miss the target and fail to fixate on it. Recognition error are the trials in which observers fixate the target but fail to identify correctly identify it. Figure 3.12 shows search and recognition errors averaged across observers. In intact and bodiless conditions the search errors dominate recognition errors. The results also show that the head and face features are important not only for recognition but also to guide eye movement during search. Search errors also increased significantly when the head and faces were removed.

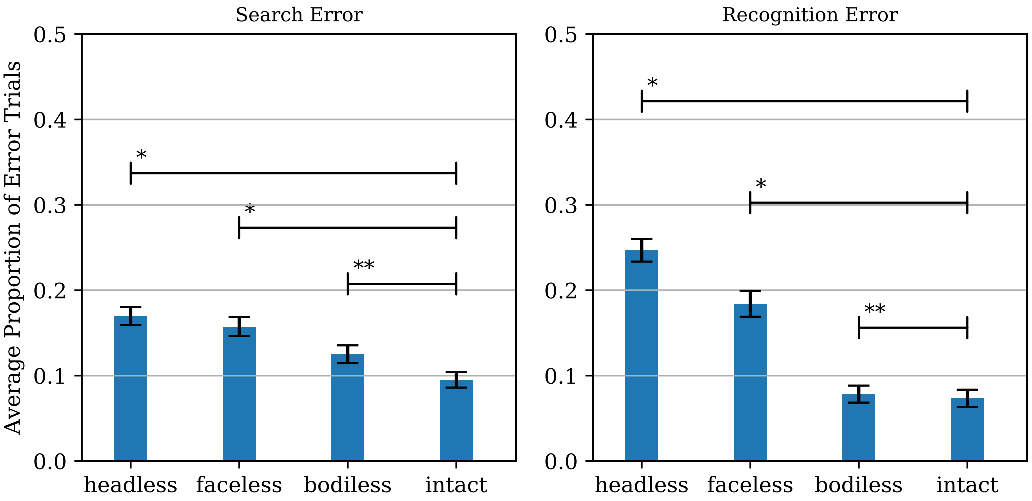


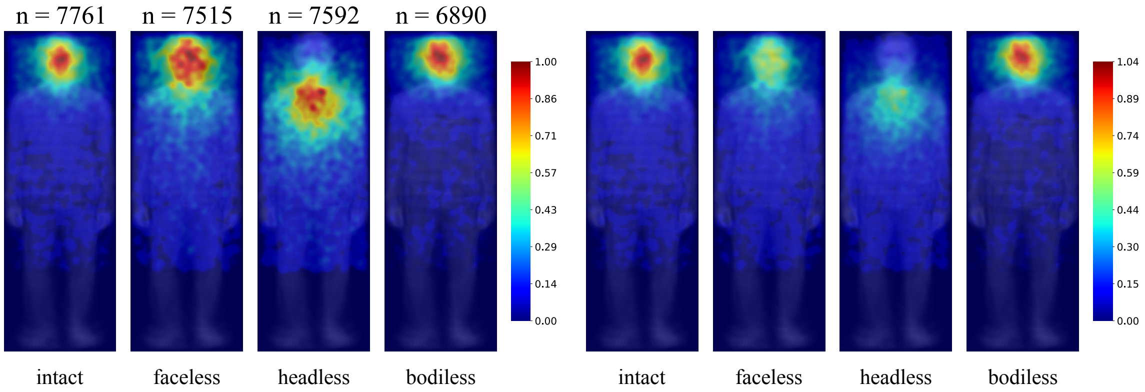
Figure 3.12: Search and Recognition Errors. The horizontal bar show the comparison between two conditions where * indicates statistical significance whereas ** does indicates not statistical significance. In intact and bodiless conditions the search errors dominate recognition errors. When the head or face is absent, the recognition errors jump more than search errors. This suggests face and head have more influence on recognition.

The averages of quantities are calculated per subject and shown in Figure 3.12. Average recognition errors are 24.67 ± 1.3, 18.41 ± 1.5, 7.83 ± 0.99 7.33±1.03 for headless, faceless, bodiless and intact conditions correspondingly.

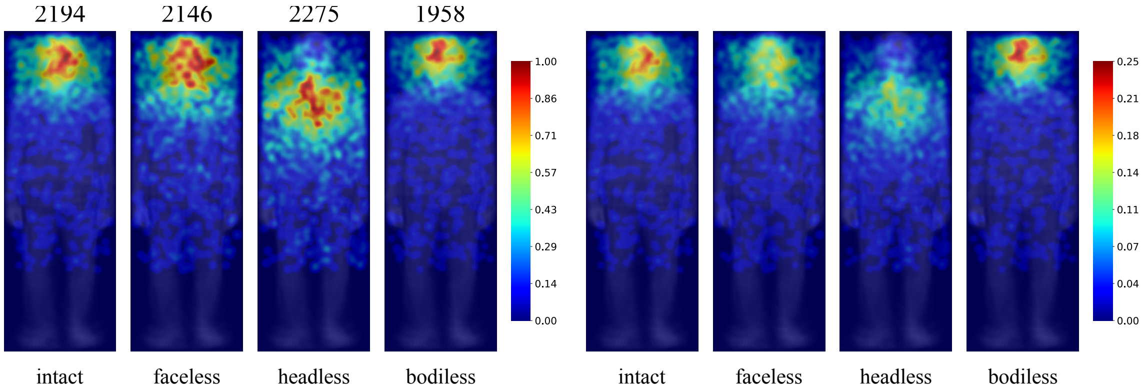
Average search errors 17.0±1.05, 15.75±1.11, 12.5±1.05, 9.5±0.9 for headless, faceless, bodiless and intact conditions. And the average errors 41.67 ± 1.78, 34.17 ± 2.2, 20.5 ± 1.6, 16.83 ± 1.5 for headless, faceless, bodiless and intact conditions. p-values of two-sided t-test comparing recognition error during headless, faceless and bodiless condition against intact condition are 1.3×10^{-09}, 3.9×10^{-06} and 0.365. Similarly p-values for search error are 1.6×10^{-05}, 1.7×10^{-04} and 0.02. Statistically significant tests are denoted by $*$ in Figure 3.12.

3.3.4 Preferred Points of Fixations within the silhouette of a person

We also analyzed the preferred points of fixation on the silhouette of the people in the videos. We analyzed a total of 20,000 fixations. To analyze the preferred points of fixations, we utilized the annotations of the people in the videos to map fixations into a normalized silhouette. Figure 3.13a and Figure 3.13b show the heat maps of the fixations across the four feature conditions. In general, we see a remarkable consistency of looking at the face of a person. The only departure from this strategy is the headless condition in which the observers look at the upper part of the body just below the location where the head would appear.

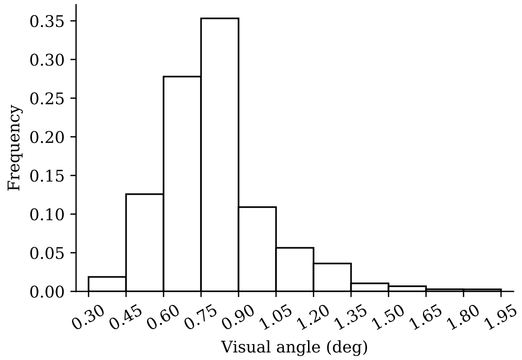


(a) All fixation densities.

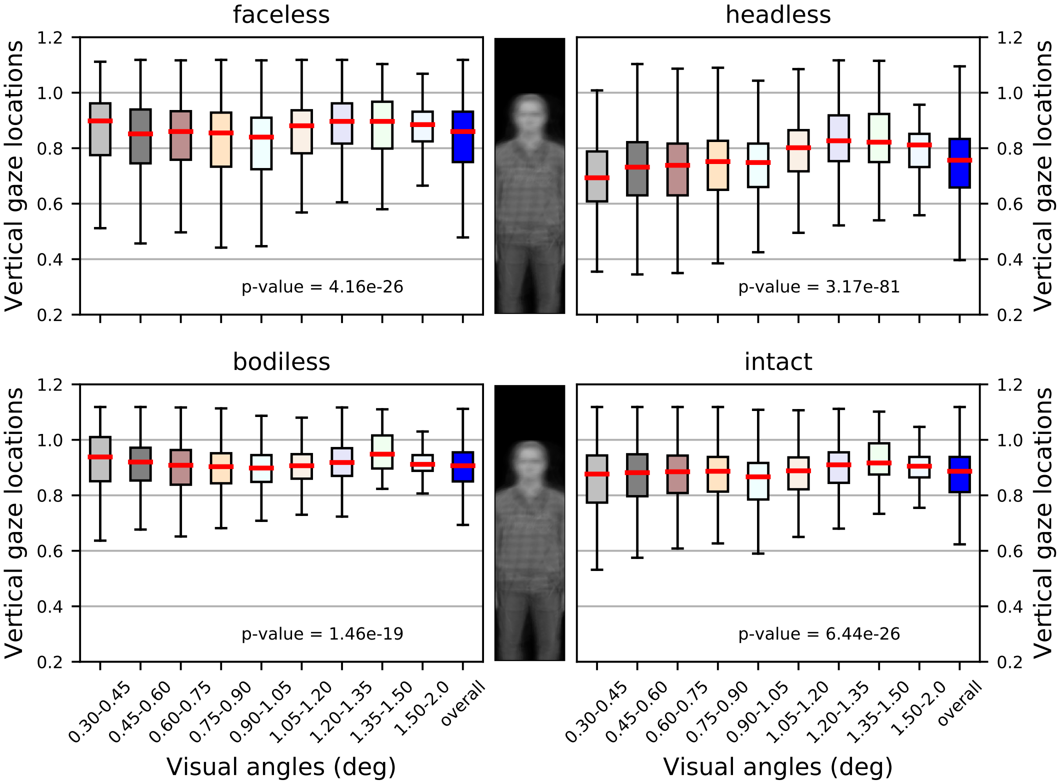


(b) First fixation densities.

Figure 3.13: Fixation distributions overlaid on a normalized person template. Left: densities are normalized within conditions, highlighting conditional distributions. Right: densities are normalized across conditions, showing relative distributions. Overall we see a strong attention bias to face and head. Especially in the absence of the face, the fixation distribution is still centered around the face. In the absence of the head, the distribution center shifts to below the face region.



(a) Distribution of face sizes.



(b) Vertical fixation distributions across face sizes. Within each condition, the fixation distributions does not change a lot when face sizes or distances to the camera change.

Because the distance of individuals from the camera varied across and within the video, the visual angle subtended by a person varied. We assessed the influence of the visual angle subtended by the individuals on the video (see methods) on fixations. Figure 3.14b shows the fixation location expressed in terms of % of distance between the top of the head and the feet for different visual angles subtended by the faces (x-axis). For all feature conditions we found statistical significant changes in the preferred points of fixation (ANOVA, p-values 4.16×10^{-26}, 3.17×10^{-81} 1.16×10^{-19} and 6.44×10^{-26} for faceless, headless, bodiless and intact conditions correspondingly). However, the magnitude of the changes with viewing angle were not large with the largest (headless condition) being a change in 10 % of the distance between the head and the feet of the indviduals in the videos. Fixation distributions on the normalized person template across different face sizes also have similar trends, as shown in Figure 3.17 and Figure 3.18.

To evaluate whether these preferred points of gaze had functional importance, we conducted a second experiment that forced observers to maintain one of four fixation locations along the midline of the person body while the stimulus was displayed for 200 ms (Figure 3.6).

3.3.5 Human Accuracy vs. FIO, CNN, FIO-CNN

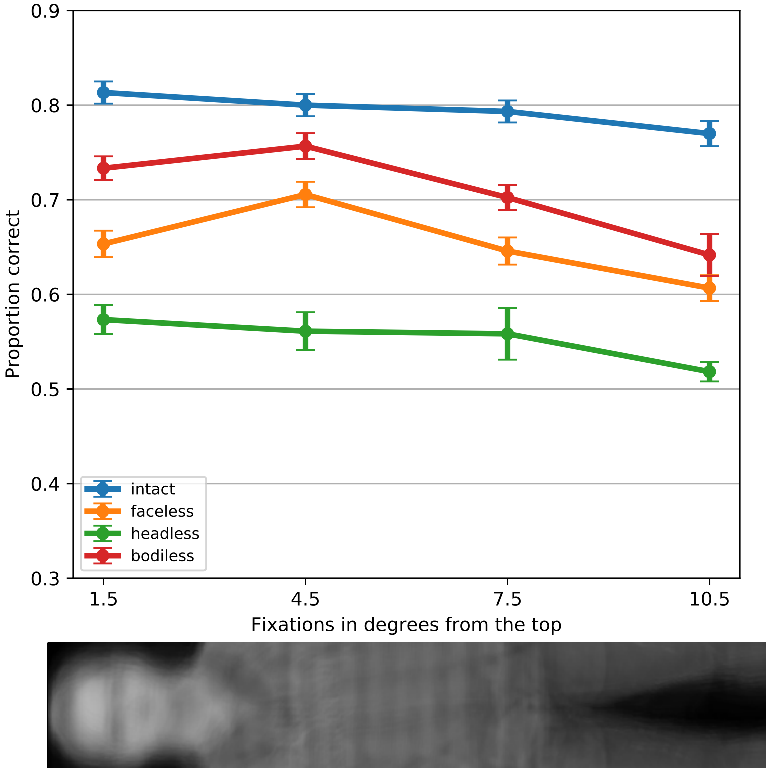


Figure 3.15: When being forced to fixate at faces subjects do not obtain maximum performance for faceless and bodiless conditions. This suggests that face-centered strategy in previous experiment does not necessarily maximize performance in human subject trials.

Fig 3.15 shows proportion correct for human observers as a function point of fixation for the four feature conditions. Detailed comparisons of different fixations within each condition is shown in Table 3.1. Except the faceless and headless conditions, fixating away from the face location led to appreciable performance degradation in terms of PC. The behavioral results show that humans guide eye movements to locations on the face that lead to high perceptual accuracy. However, these results do not necessarily show that humans enact gaze patterns that are optimized for the statistical distribution of discriminating information present in the human face combined with the foveated nature of the human visual system.

distance to top (deg)	1.5	4.5	7.5	10.5
intact	-1	0.9914	0.8188	0.0189
headless	-1	0.63	0.63	4.2×10^{-3}
faceless	9.5×10^{-3}	-1	3.6×10^{-3}	2.9×10^{-6}
bodiless	0.21	-1	5.9×10^{-3}	4.7×10^{-5}

Table 3.1: For each condition, the fixation with peak human performance is chosen as the anchor to compare against performance at other fixations. The table show p-values for two-sided t-test of such comparisons.

We investigate the model performance compared to humans. Different from humans, the FIO, CNN and FIO-CNN performance degraded as much or more for the bodiless condition. At the same time, they do not degraded as much as human in the absence of faces or heads.

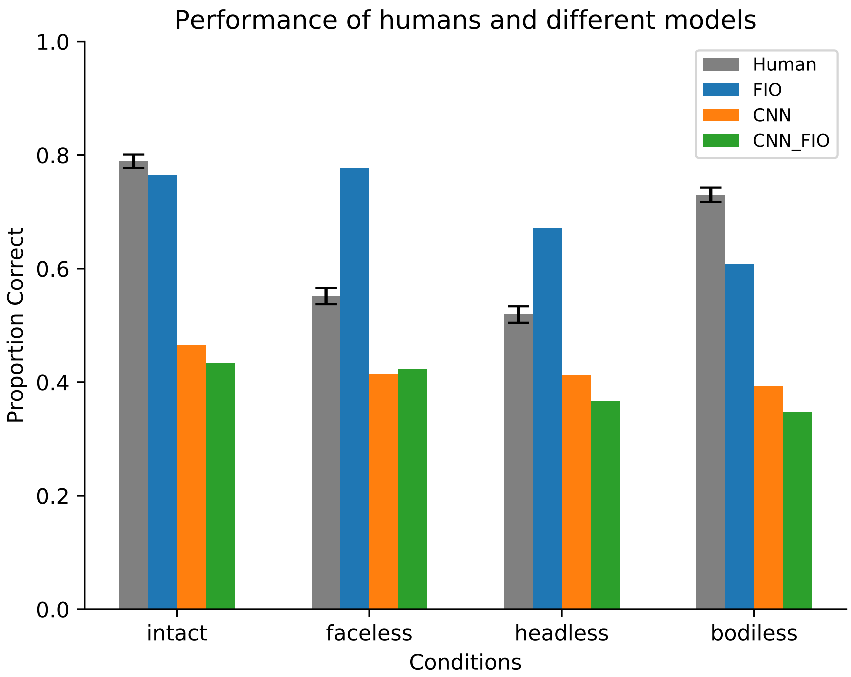


Figure 3.16: Human and models performances (fixate at faces). Models' performances degrade as much as or more than humans for the bodiless condition. At the same time, they do not degraded as much as human in the absent of faces or heads.

3.4 Discussion

In static scene, it is commonly believed and has been shown that humans tend to shift their attention toward regions that are most useful for recognizing or identifying people (for maximal information gain). In most cases they tend to look at the face when it is available and look at the body when the face information is not useful. Our study shows that, in unconstrained dynamic scene, especially viewing time is limited, humans tend to look at the face region when its information is not even available. Our forced fixation experiments and FIO models also suggest maximal information location should be shifted below the face. This demonstrates the face-centered human strategy is not related to higher information content in faces, but likely arises as a byproduct of the importance of faces for other evolutionary critical tasks.

The fact that the popular machine models fail to weight the importance of faces for the task and their performance fall behind humans by a large margin suggests the need for modeling human eye movements. We will explore this further in the next chapter.

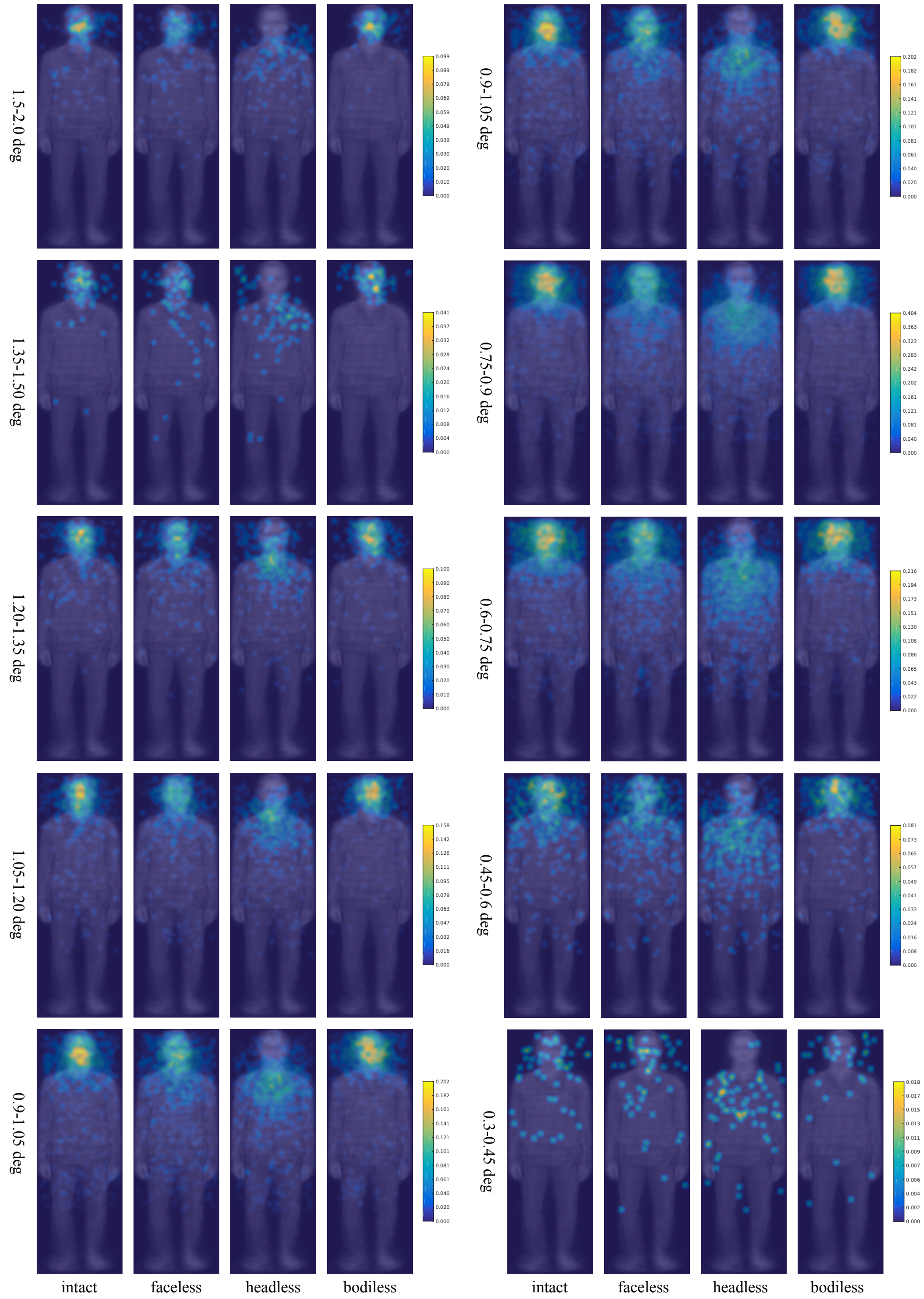


Figure 3.17: Fixation distributions at different face size.

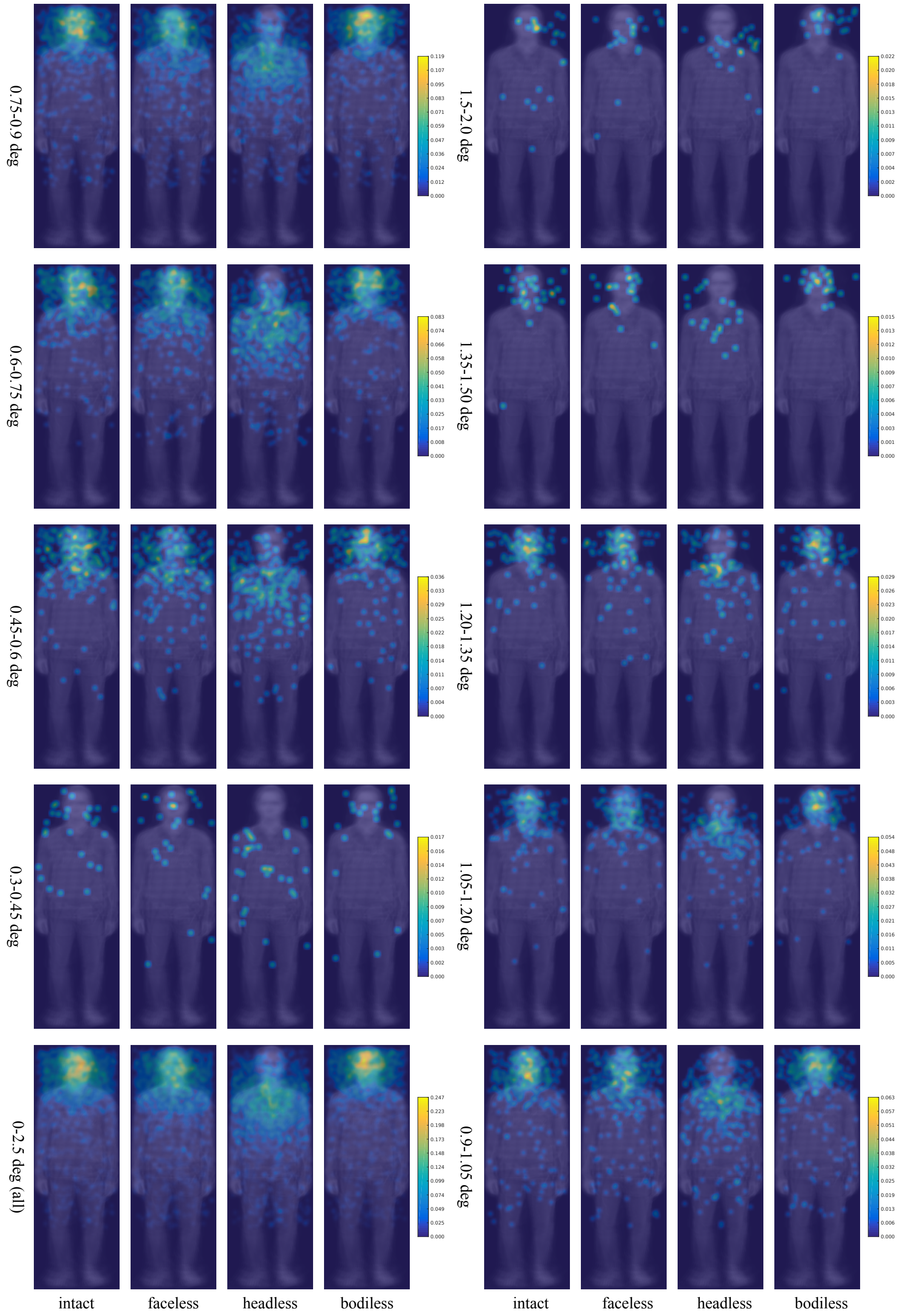


Figure 3.18: First fixation distributions at different face size.

Chapter 4

Predicting fixation points with a recurrent neural network

Chapter 3 shows that humans make distinct eye movements towards the faces while searching for other people. It also shows that machine models failed to weight the important of facial features for the task. Motivated by these findings we aim to model how humans make eye movements or gaze. Given one image, the model predicts what would be the sequence of locations a human would likely look at in the free viewing task. Such a model could be classified as a visual attention model. Recent works in visual attention, however, only focus on saliency map prediction, discarding the sequential aspect of human eye movements. Since the order of exploration is important when encountering new environments, the ability to make gaze prediction is needed, particularly for active vision applications.

In the following, We present a model that generates close-to-human gaze sequences for a given image in the free viewing task. The proposed approach leverages recent advances in image recognition using convolutional neural networks and sequence modeling with recurrent neural networks. Feature maps from convolutional neural networks are used as inputs to a recurrent neural network. The recurrent neural network acts like a visual working memory that integrates the scene information and outputs a sequence of saccades. The model is trained end-to-end with real-world human eye-tracking data using back propagation and adaptive stochastic gradient descent. Overall, the proposed model is simple compared to the state-of-the-art methods while offering better performance on a standard eye-tracking data set.

4.1 Introduction

Due to sensory and computational limitations, humans and many other animals employ visual attention as the strategy to actively explore the environment. Human visual system only has high visual acuity in a small region, the fovea, and the photoreceptor density drops rapidly when moving away from the fovea [54]. When a human observer gazes at a point, the fixated region is projected onto the fovea and sampled with highest density. The peripheral on the other hand is perceived with low resolution. This helps reduce the amount of information the brain needs to process at a given time but it requires the eyes to move constantly to integrate the information of the entire scene. The mechanisms which control such eye movements have been extensively studied in psychology and neuroscience [55, 56].

A similar computational bottleneck exists in computer vision where processing the entire image might be prohibitively expensive. For example, the popular deformable part model for object detection [57] takes a few seconds to process a single image as it uses scanning windows over the whole image. In face recognition or image classification with convolutional neural networks, the

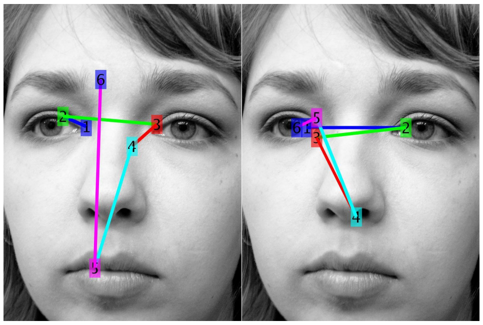


Figure 4.1: Given an image we would like to predict a sequence of fixations that a human might look at. On the right is the image with fixation sequence predited by our model. The image on the left is actual eye tracking data from mutiple subjects.

input image normally needs to be cropped so that objects are aligned roughly at the center of the image. It could be advantageous in these cases to have an attentional model to select meaningful regions to process. Toward this, recent work in computer vision has focused on (1) saliency models, which predict the probability map of fixations; and (2) objectness measures, where potential regions containing objects are selected. However, these models ignore the sequential nature of visual attention, which could be valuable information for visual search and large scale image analysis.

Predicting the fixation sequence is quite challenging and has not received much attention in computer vision. Relevant work include [58, 59], however they do not use the temporal information in the eye-tracking data. In [60, 61] the models tend to be specific to the datasets of interest. In contrast, the proposed model is simple and does not require either prior information or feature-engineering. Our model takes advantage of recent advances in image recognition with convolutional neural network and sequence modeling with recurrent neural network to achieve comparable performance to the state of the art methods.

4.2 Related Work

Since saliency is an important block in visual attention modeling, we first review existing saliency models. Following this, related works on gaze scanpath predictions and sequential modeling are discussed.

4.2.1 Saliency Models

A comprehensive review of saliency models can be found in [62].

Cognitive methods propose biologically plausible computational architectures to compute saliency. In [7] center-surround feature maps are computed from the Gaussian pyramid for color, intensity and orientation channels, and then combined into a single saliency map. Graph-Based Visual Saliency (GBVS) model [8] extracts similar features as [7] but builds a graph associated with each feature map. Saliency maps are then the stationary distributions of a markov chain induced by the graphs, and finally combined into a single saliency map. The model proposed in [9] is based on more features of human visual system including center-surround property, contrast sensitivity function, visual masking and perceptual decomposition.

Information theoretic methods select the most informative regions as salient. In Attention based on Information Maximization (AIM) model [63], Shannon's self-information measure is used to compute saliency. Saliency Using Natural statistics (SUN) model [64] gathers natural image statistics which is subsequently used to compute the difference against current image statistics as a new kind of self-information. In [65] saliency is modeled by minimizing the conditional entropy of a local region given its surroundings. In [66] authors use Incremental Coding Length (ICL) to measure the perspective entropy gain of each visual feature and select features with maximinum coding length increments for saliency.

Learning-based methods learn models from recorded eye-tracking data or labeled saliency regions. Discriminative features are extracted from each image location to compute the saliency probability. These features could contain high level semantics (e.g., faces or text) and therefore could be used to model the top-down attention. This group of methods includes task dependent attention models [67, 68, 69], and saliency models based on conditional random field [70, 71], support vector machine [72, 73], and convolutional neural networks (CNNs) [74, 75]. These learning based models could be useful for gaze sequence predictions [61].

4.2.2 Scanpath Models

Scanpath models aim to predict an ordered set of fixations for a given image. The first model [7] generates a scanpath from a static saliency map.

The model uses a Winner-Take-All neural network and inhibition of return scheme to output a sequence of winners as the fixation predictions. In [76] the authors introduce a stochastic model for scanpath generation. They show that distribution of natural scanpath magnitudes is similar to Levy distribution , which has a power law dependency in the saccades magnitudes. This approach is extended in [77] to model eye gaze shifts using Levy flight, a random walk in which the step-lengths follow Levy distribution, but each jump has an acceptance probability determined by gain of saliency. [59] proposes a model to generate scanpaths on natural images based on the principle of information maximization. The model exploits three factors guiding sequential eye movements: reference sensory responses, fovea periphery resolution discrepancy, and visual working memory. [60] introduces the first learning based method for gaze sequence prediction by integrating semantic information along with Levy flight and saliency map into a Hidden Markov Model. In [61] authors use reinforcement learning to learn a fixation policy to obtain state of the art performance.

4.2.3 Sequence Modeling

Learning-based methods have achieved better performance for fixation sequence predictions since they utilize the ordered information available in the

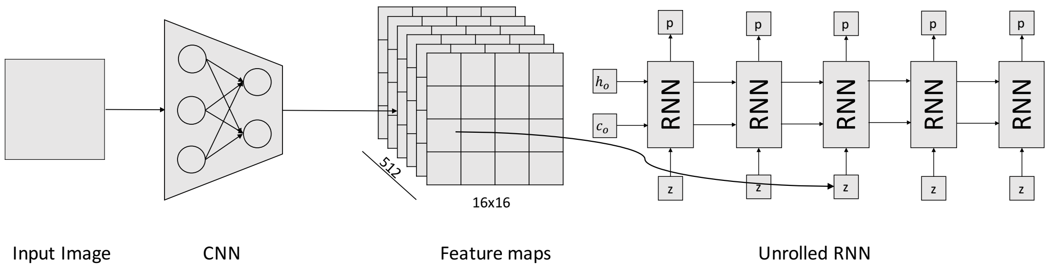


Figure 4.2: The overview of our model. Each location (a bin among 16x16 bins) in the input is mapped to a feature vector of dimension 512 in the feature maps. Given the feature at current location, the RNN outputs the distributions of next locations.

training data. Recurrent Neural Networks (RNNs) have recently shown to be an elegant and flexible approach to process sequences, either as input, output or both. While RNNs are hard to train, some versions can be trained effectively and obtain the state of the art performance in several sequence prediction problems including machine translation [78], image captioning [79] and video description [80].

In this work, we utilize both CNN and RNN's power and simplicity to model fixation sequences. Compared to the state of the art [61] and other previous works [60, 62], our model offers several advantages. First, it does not assume any prior knowledge about the data. Most other works integrate some understanding about eye tracking data in the models. For example, the center bias and Levy flight distribution of eye movements have been used either as features or priors in [61, 60, 62]. The model in [61] is even trained with different set of features for different eye tracking datasets. Due to end-to-end training, we expect our model to learn such knowledge from the data itself without any feature-engineering efforts. Second, the proposed method unifies the feature extraction steps into a single pass through a CNN. Other methods normally need to extract low level features such as edges or color features and then compute semantic features using object detectors.

4.3 Approach

The overall approach is shown in Figure 4.2. Given an image, we first extract features from a pretrained CNN. These features are then used as inputs into a RNN to make predictions about the fixation locations. We now will discuss in detail each component of our model.

4.3.1 CNN Feature extraction

Many recent works exploited transfer learning or domain adaptation using pretrained CNNs [79, 81]. The idea is that knowledge gained from training millions of images for classification could be used for different tasks. The pretrained features from convolutional neural network have been shown to be effective for many different tasks ranging from fine-grained image classification [82], image segmentation [81] to image caption generation [79]. These features contain not only the semantic information of the images as a whole, but also the locations of such information in images [83] which make it an appropriate choice for our model. We modify the original 16-layer VGG network [84] to retain only convolutional layers of the network. This helps us maintain the spatial information in the extracted feature maps and allows to run the network with different image sizes. In our experiments, we use input size of 512x512 and the resulting output feature maps are of size 16x16. Since we only use convolution and pooling layers these maps represent 256 regions in the input image. We modify the original 16-layer VGG network [84] to retain only convolutional layers of the network. This helps us maintain the spatial information in the extracted feature maps and allows to run the network with different image sizes. In our experiments, we use input size of 512x512 and the resulting output feature maps are of size 16x16. Since we only use convolution and pooling layers these maps represent 256 regions in the input image.

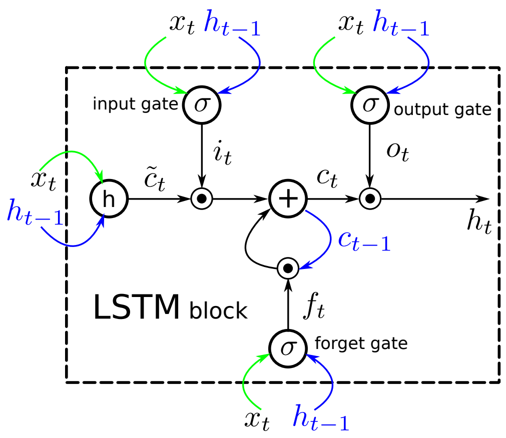


Figure 4.3: Illustration of a LSTM block. The green connections are feedforward with associated weight matrices. The blue are feedback weight connections.

4.3.2 Spatial Quantization

In line with the CNN feature extraction, we spatially quantize the input image into 256 regions by a 16x16 grid. Each region could be represented by a 512-dimensional feature vector from the CNN feature maps. At the same time eye tracking fixation data are binned into those regions. Each fixation is now represented by the center of the region it is in, and a sequence of fixations is a sequence of jumps from one region to another. This greatly simplifies the modeling process compared to other models where fixation locations are at super-pixel level [60, 61]. In line with the CNN feature extraction, we spatially quantize the input image into 256 regions by a 16x16 grid. Each region could be represented by a 512-dimensional feature vector from the CNN feature maps. At the same time eye tracking fixation data are binned into those regions. Each fixation is now represented by the center of the region it is in, and a sequence of fixations is a sequence of jumps from one region to another. This greatly simplifies the modeling process compared to other models where fixation locations are at super-pixel level [60, 61].

4.3.3 Long Short Term Memory (LSTM)

LSTM [85] is designed to mitigate the vanishing and exploding gradients during training of recurrent network, and has been widely used for sequential modeling. Training a traditional RNN could be difficult because the gradient signal is multiplied many times by the recurrent weight matrix during back propagation. If the weights are small, the resulting gradient signal could be so small that learning will become either too slow or even stop working (vanishing gradients). On the other hand, if the weights are large, the gradient signal could end up being too big and cause the learning to diverge (exploding gradients). LSTM aims to fix these issues by introducing a modular structure, referred to as a memory cell, as in Figure 4.3. A memory cell includes an input gate, an output gate, a forget gate and a self-recurrent connection. The gates control how the cell modulate its dynamics and its interactions with the input and the output. The forget gate modulates whether the cell should remember or forget its previous state. The input gate controls how much of the input would have an effect in the cell, and the output gate controls the effect of the cell at the output (on other neurons).

An example of a LSTM block is shown in Figure 4.3 and its main computations with intput x_t, output h_t and its recurrent cell c_t are as follows:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{4.1}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{4.2}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{4.3}$$

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{4.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4.5}$$

$$h_t = o_t \odot c_t \tag{4.6}$$

Here the gates $\{i_t, f_t, o_t\}$ are modeled as perceptron units using sigmoid activation σ. $\{W_i, W_f, W_o, W_c\}$ and $\{U_i, U_f, U_o, U_c\}$ are the weight matrices for the feedforward and feedback connection correspondingly; $\{b_i, b_f, b_o, b_c\}$ are biases. The gates' modulation operator $\odot$ is element-wise multiplication.

4.3.4 Scanpath Model with LSTM

The key observation is that saccade planning is not memoryless, i.e. it is influenced by the gaze history [86]. Such visual working memory could be naturally modeled using a LSTM. In our case, the LSTM models the conditional transition probability of the next fixation given the current fixation and the history of information it has seen so far. The history is expected to be remembered in the state vectors c_t and h_t. The transition probability p_t is computed as follows:

$$z_t = \mathrm{CNN}(S_t) \tag{4.7}$$

$$x_t = \mathrm{FF}_{\mathrm{x}}(z_t) \tag{4.8}$$

$$h_t = \mathrm{LSTM}(x_t, h_{t-1}, c_{t-1}) \tag{4.9}$$

$$y_t = \mathrm{FF}_{\mathrm{y}}(h_t) \tag{4.10}$$

$$p_t = \mathrm{softmax}(y_t) \tag{4.11}$$

Here we use CNN to extract the feature z_t at the current location S_t. FFs are one-layer perceptrons: the first one is used to map the dimension of z_t to the number of LSTM cells and the second is used to map LSTM output to the distribution over possible locations:

$$P(S_{t+1}|S_t, c_{t-1}, h_{t-1}) = p_t \tag{4.12}$$

In order to predict the first fixation, the initial states of LSTM is also computed from the CNN features using one-layer perceptrons:

$$h_o = \mathrm{FF}_{\mathrm{h}}(\mathrm{CNN}), c_o = \mathrm{FF}_{\mathrm{c}}(\mathrm{CNN}) \tag{4.13}$$

4.3.5 Learning

We aim to estimate parameters θ of the model (all weight matrices and biases) from eye tracking data. The log-likelihood of one fixation sequence $S = \{S_2, S_3, \ldots, S_{n+1}\}$ given an image I:

$$L(S|I;\theta) = \sum_{t=1}^{n} \log P(S_{t+1}|S_t, c_{t-1}, h_{t-1}) \tag{4.14}$$

$$= \sum_{t=1}^{n} \log p_t(S_t) \tag{4.15}$$

The model is optimized so that the log-likelihood over all training samples is maximized:

$$\theta^\star = \arg\max_{\theta} \sum_{(I,S)} L(S|I;\theta) \tag{4.16}$$

The equivalent form of this is the minimization problem on traditional log loss (negative log-likelihood). From Figure 4.2 we can see that the computation flow of equation 4.15 for each sequence is actually a DAG and thus the learning can be performed using standard back propagation.

4.3.6 Sequence Prediction

There are multiple approaches that can be used to generate a sequence from the trained model. The simplest prediction scheme is to sample the most probable location given the current location S_t and the states of the LSTM:

$$L_{t+1} = \underset{S_{t+1}}{\arg\max}\, P(S_{t+1}|S_t, c_{t-1}, h_{t-1}) \tag{4.17}$$

We sample the first location according to p_1. Given the feature at that location we can sample from p_2, and continue until we reach some maximum pre-defined sequence length.

The optimal way would be searching for the sequence with maximum likelihood based on equation 4.15. However, it is too expensive because of the exponential growth in the number of sequences. Instead we use beam search to generate m best sequences with largest likelihoods. We do this by always maintaining m best candidate sequences at each step. At the end of the step, each candidate will have 256 children for the following locations. Among all resulting sequences we choose m of them with the maximum likelihood. Finally, we return the best among m candidates as the predicted sequence. We use the beam search in the our experiments with a beam of size 20.

We can also simulate the stochastic scanpaths by considering $P(S_{t+1}|S_t, c_{t-1}, h_{t-1})$ as a multinomial distribution and sample the next fixation from this distribution.

4.4 Experiments

4.4.1 Evaluation metrics

There are several metrics have been proposed to measure the consistency of eye tracking sequences. In [59] authors employ time-delay embedding [87]. This method divides each sequence into segments of length k, starting at some order t. By varying t, we have a set of vectors representing the sequence. The similar between two sequences is then measure by the distance between the two represented sets. However, multiple segments of a sequence might be matched to the same segment of another sequence. Recently [88] proposes a method based on the Dynamic Time Warp algorithm (DTW) [89]. The algorithm represents each scanpath as a geometric vector and calculates similarity using some geometric measure. A more popular approach is to use string alignment algorithms [90]. The main idea is to think of sequence as string and the distance between two sequences is then the cost it takes to align so that the two strings are matched. The notion of match could be different. For example in [60], two fixations are matched if they are within a certain spatial distance. We use similar approach as [61]. First all fixations are clustered and fixations in each cluster will be assigned a unqiue label (alphabet). The sequence is then represented as a string of alphabets and two fixations are matched if two associated alphabets are the same. We use Needleman-Wunsch string matching algorithm to compute distances [91]. A predicted sequence is compared to data from all of the human subjects and scores are averaged to obtain the final score. We also compute score between any two subjects and average them to get the upper bound for the performance.

Similar to [61], we use meanshift clustering to assign labels for fixations. The meanshift bandwidths are chosen to maximize interaction among clusters:

$$I = \frac{N_b - N_w}{C} \tag{4.18}$$

Here N_b is number of fixation transition between clusters, N_w is number of transitions within clusters and C is the number of clusters.

4.4.2 Dataset

We evaluate our model using the MIT [72] dataset. The dataset contains 1003 images with various types of object categories and scenes (indoor, outdoor, landscape and portrait). Eye tracking data was collected during a free viewing task with 15 subjects per image, with a total of about 15,000 sequences. This is currently the largest free-viewing eye-tracking dataset avail-

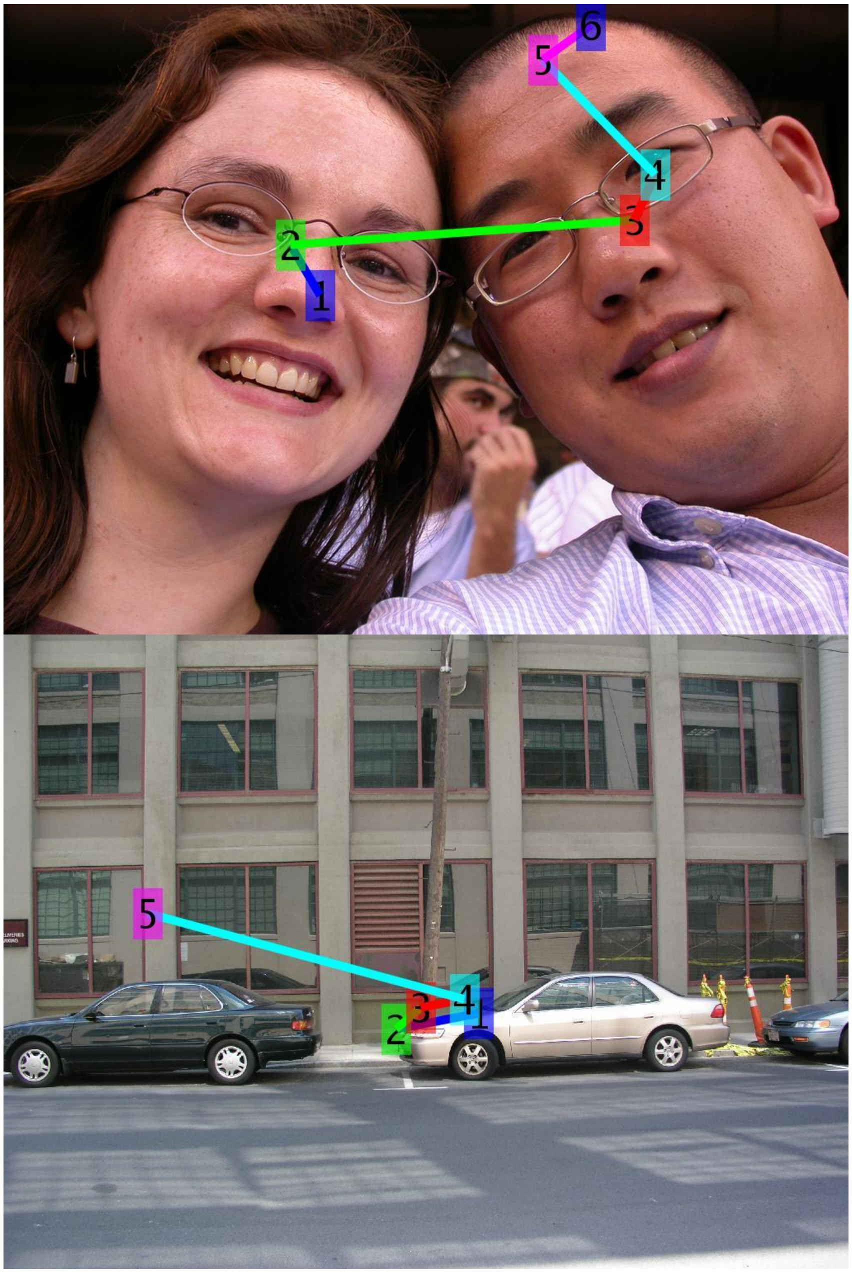


Figure 4.4: Sampled images in MIT1003 dataset with eye tracking data overlaid.

able for natural images. Sampled images from the dataset are shown in Figure 4.4.

4.4.3 Training Details

Due to limited training data, the main challenge is to deal with overfitting. To reduce overfitting, we only train the RNN part of the model, leaving the CNN part untouched (no fine-tuning). We do a simple data augmentation with horizontal flipping: each image and its eye-tracking data are flipped to create more training examples. Even though eye-tracking data are not absolutely invariant to this transformation (humans might not look at flipped texts for example), we obverse slight improvement in the results. We also try to keep the number of model paratmeters small, limiting LSTM size to only 64 recurrent dimensions. The model is trained with dropouts [92] using rmsprop [93], a variant of adaptive stochastic gradient descent, without using momentum term. The LSTM weights are initialized with random orthogonal matrices, and the remaining weights are randomly initialized with a small variance. The model is developed and trained with the Theano [94] library.

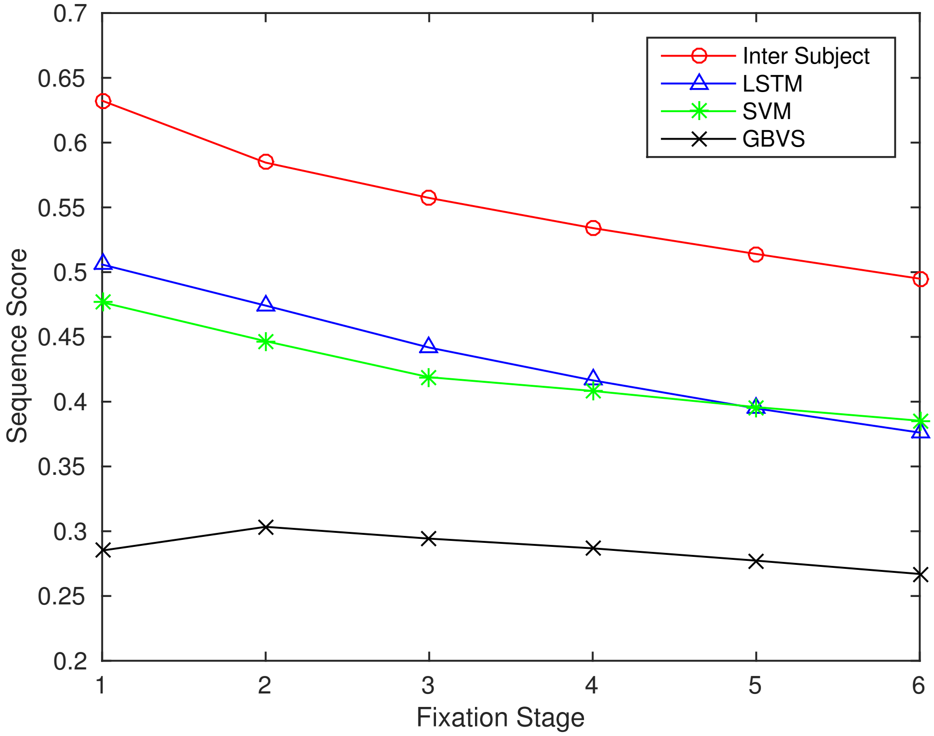


Figure 4.5: Evaluation of the RNN model (blue triangle) and baseline models GBVS [8] (black cross) and Judd [72] (green star) and inter-subject performance (red circle) on the MIT dataset.

4.4.4 Results

It was first shown in [62] that when using winner take all method (WTA) to compute fixation sequences, GBVS [8] and Judd [72] saliencies perform significantly better than other saliency models. Recently [61] has shown that SVM saliency models like Judd's with WTA fixations can perform comparably

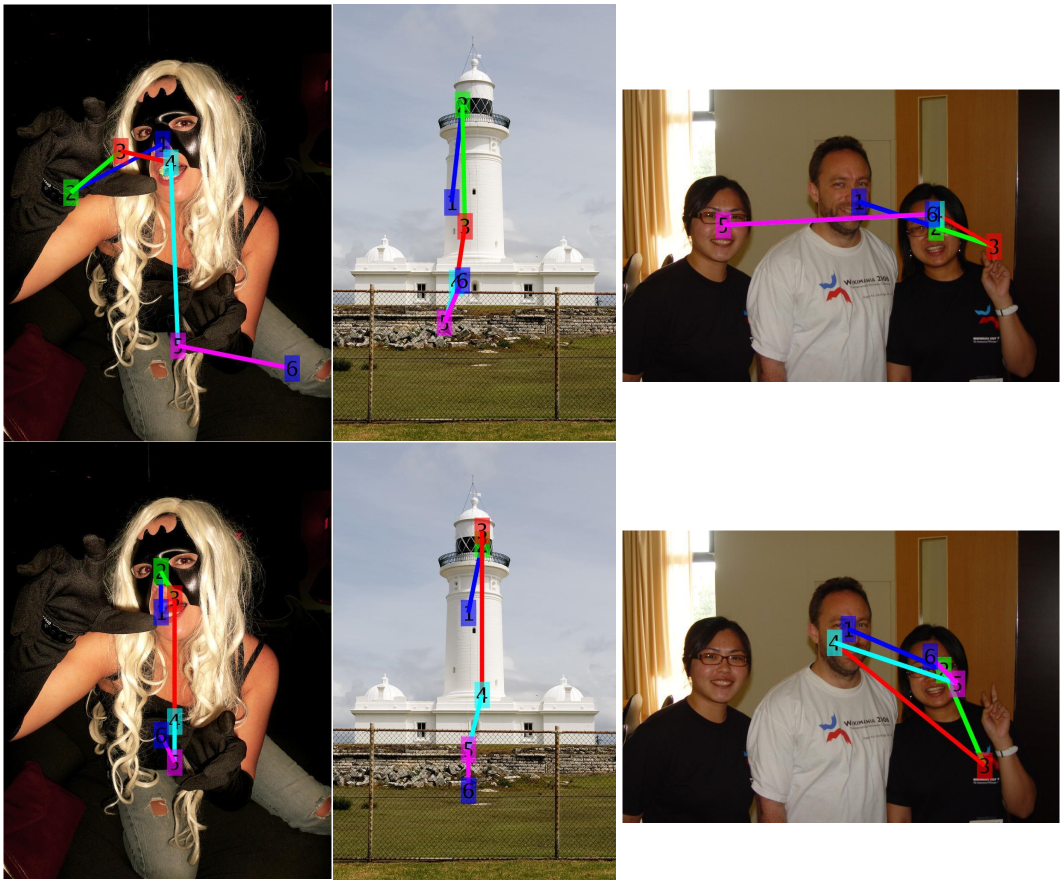


Figure 4.6: Top row: human fixation sequences. Bottom row: predicted sequences. The right most image examples illustrates a failure case where one of the faces is completely missed by the RNN model. Only the first few fixation points are shown. The numbers inside the square nodes correspond to the sequential order of these points.

with state of the art [61] itself. Since we were not able to get access to the evaluation codes of the referenced models [59, 60, 61], we could only compare our models with WTA using Judd and GBVS saliencies. We randomly split the data, using 90% for training and 10% for testing. We retrain the SVM saliency model using the same training data. The LSTM model uses beam search with beam size 20. The comparisons are limited to 6 fixations for each sequence (similar to [61]). Figure 4.5 shows the performance of three models on the test set. The inter-subject performance is calculated over the whole dataset. Our method performs slightly better than WTA on SVM (Judd's saliency).

Figure 4.6 shows some sample results from our model. Comparing with the human-observed fixation sequences, we see that the model predictions correspond to meaningful semantic regions.

4.5 Conclusion

We presented a model to predict a sequence of fixations that humans are likely to look at in a given image (in free-viewing task). We present a simple framework to model sequences with recurrent neural networks using localized features extracted from a pre-trained convolutional neural network. The model is trained to maximize the likelihood of fixation sequence given an image using free-viewing human eye-tracking data. Despite its simplicity and limited data, we achieve favorable performance compared to more complicated methods. With recent advances in eye-tracking technology, we would expect better performance of our model when large eye-tracking datasets become available in the future. Exploring a similar framework to predict dynamic gaze in videos will have interesting applications in video object tracking and human assisted annotations of large data sets. We have presented a model to predict a sequence of fixations that humans are likely to look at in a given image (in free-viewing task). We present a simple framework to model sequences with recurrent neural networks using localized features extracted from a pre-trained convolutional neural network. The model is trained to maximize the likelihood of fixation sequence given an image using free-viewing human eye-tracking data. Despite its simplicity and limited data, we achieve favorable performance compared to more complicated methods. With recent advances in eye-tracking technology, we would expect better performance of our model when large eye-tracking datasets become available in the future. Exploring a similar framework to predict dynamic gaze in videos will have interesting applications in video object tracking and human assisted annotations of large data sets.

Chapter 5

Foveated Image Recognition

In this chapter we introduce a new convolutional network model for working with foveated image data. Given a fixation and a foveated representation of the input around that fixation, the model predicts the likelihood of the object class it belongs to. Since foveation comes at the cost of active scanning of the scene/image data, a greedy method is proposed of potential fixation locations to explore. The model is the first of its kind that is applicable for active foveated vision. Since foveated sensors are not yet generally available, the experiments described in this Chapter use standard rectangular pixel images that are transformed through a log-polar transformation to create the foveated representations for recognition.

5.1 Introduction

In many computer vision applications, including robotic vision and drone systems, there is a need to actively monitor the environment. This requires that the sensor scan the field of view for activities or objects of interest. While most existing vision sensors sample the visual field uniformly, there is a significant potential for utilizing non-uniform sampling such as the ones found in biological foveated vision systems. The primary motivation is in the computational efficiency offered by such foveated sensors that typically required a small fraction of the computing power for feature extraction and gaze prediction. Multiple fixations will enable constructing a comprehensive description of the scene for specific tasks. As noted in earlier, previous research in this context has focused on mechanisms of eye movements (attention) [29] and neural representations of foveated sensory inputs [95]. Most research in machine vision has involved analysis of passively sampled images. *In the following we present one of the first approaches to utilizing foveated samples in the context of image recognition, and demonstrate the feasibility of simultaneously optimizing the recognition and gaze prediction tasks.*

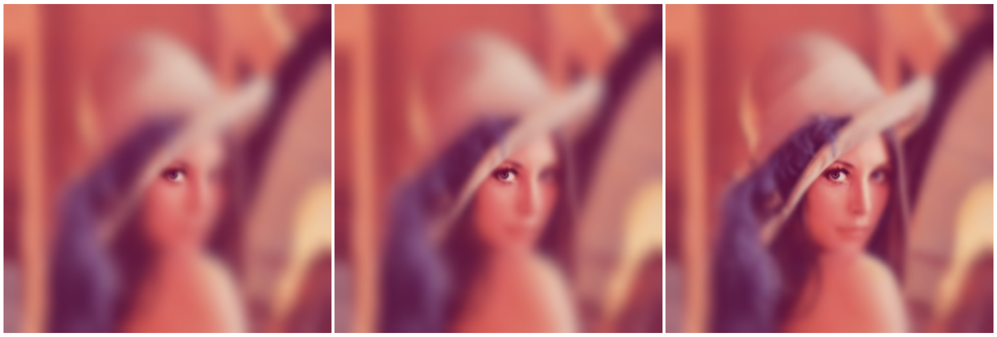


Figure 5.1: Traditional foveated images at different scales [96]. A spatially varying filter is used to obtain a foveated version of the original image at the same resolution with decaying level of blurring when moving away from the fixation.

5.2 Related Work

Recent works aiming on foveated image recognition [97, 98] use high resolution followed by appropriate spatial filtering to simulate the foveated data. In [97] a spatial filtering step is required to obtain a foveated version of the same resolution of the original image with decaying level of blurring when moving away from the fixation. An example of such filtering operations are shown in Figure 5.1. In [98] the authors first compute the histogram of gradients (HOG) features on the original images, followed by a foveated template which pools HOG features for each bin of the template, which is then used to make decision on object class or next set of fixations. These methods demonstrate the potential of foveated recognition but fail to take advantage of the reduced resolution of such representations.

Given the challenges associated with processing and integrating multiple fixations, including gaze prediction and working with non-uniformly sampled data, there are very few works on image/object recognition using foveated samples. In [99] the authors use foveated images in communication for bandwdith reduction. In [100] a corner detector was developed based on Moravec operator [101] to model overt attention of foveated vision. More recently [23] used Fourier transform to detect lines and circles in log-polar images. In the following we present one of the first convolutional neural network architectures that work directly with the foveated image inputs for image recognition. Further, the model can obtain better accuracy with more fixations, thus can be easily adapted to varying computational resources.

5.3 Foveated Representation

There are many ways to create a foveated images from regular 2-D images. Multiresolution methods maintain different versions of uniform resolution images at multiple scales. They do not require bookkeeping of sophisticated geometric transformations, and also work well with existing image processing methods. However, the representation is redundant and cannot represent images acquired directly by foveated sensors. Geometric methods use geometric transforms or look up tables to build a non-uniform sampling grid, which can be interpolated back to a uniform grid for visualization. In this work we use the log-polar transformation since it closely follow cortical mapping in in the retina The idea is to quantize the space into bins in the log-polar domain. In our experiments we use 48 angular bins and 48 radial bins, see Figure 5.2. An example of a reconstructed image from the log-polar samples is shown in Figure 5.3.

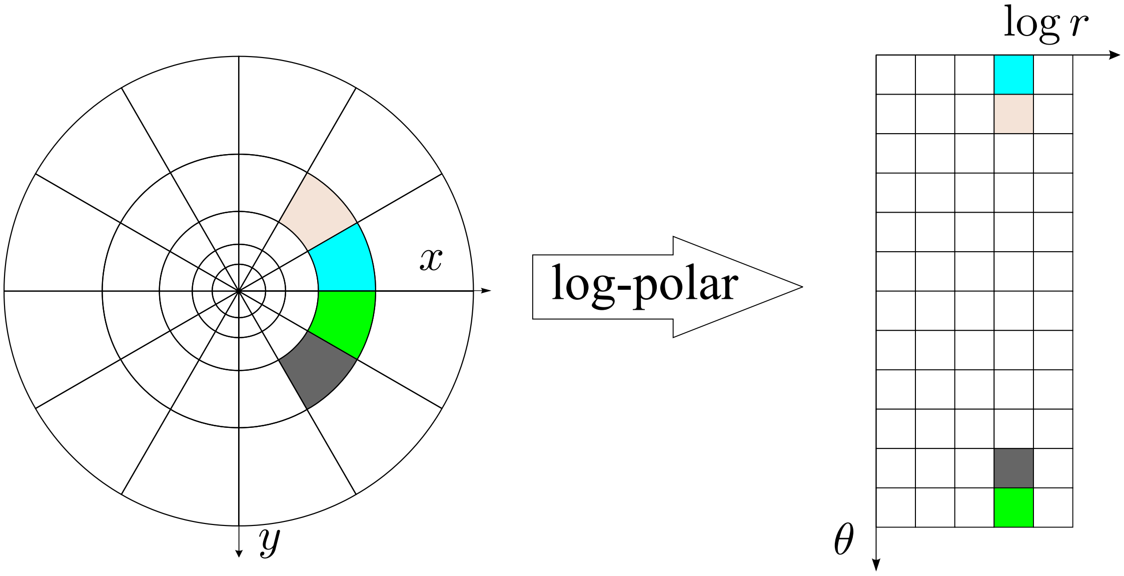
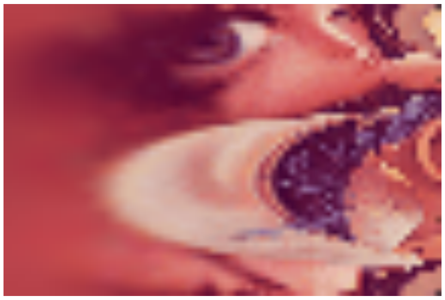
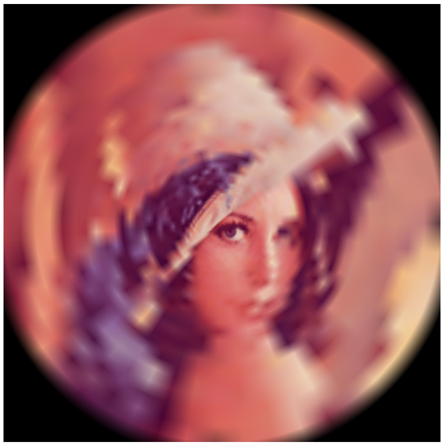


Figure 5.2: Log-polar sampling: A log-polar mapping followed by bilinear interpolation is used to map the original cartesian grid to the log-polar array.

(a) Original image at 512x512.



(b) Log-polar transformed image at 48x72 (scaled up for visualization).



(c) Image reconstructed from the log-polar data in (b) above, scaled to 512 × 512 pixels.

Figure 5.3: Log-polar transform and reconstructed image.

5.3.1 Group Convolution

In a normal convolutional network a scanning window of kernel size $h_1 \times w_1$ is correlated with corresponding locations in the input, see Figure 5.4. In contrast, in group convolution, the input and output are divided into groups and a similar operation is performed within each group. This provides a way to make separate computation streams within a network by specifying same number of groups for all convolutions. We use group convolutions throughout the network to separate convolutional operators into 4 distinct groups, each ideally would represent a processing stream at different resolutions.

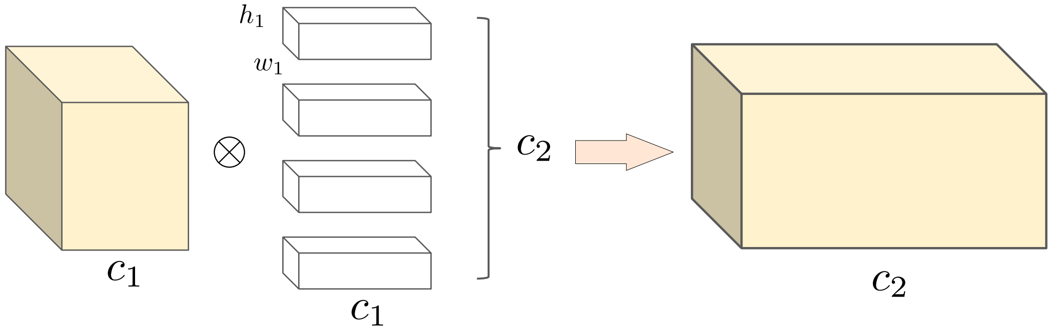


Figure 5.4: Normal convolutional network: In a conventional convolutional network, given an input featue map of dimension c_1, output feature map dimension c_2, and filters of size $h_1 \times w_1$, the output feature map is computed as the dot product of the filters with the corresponding spatial input.

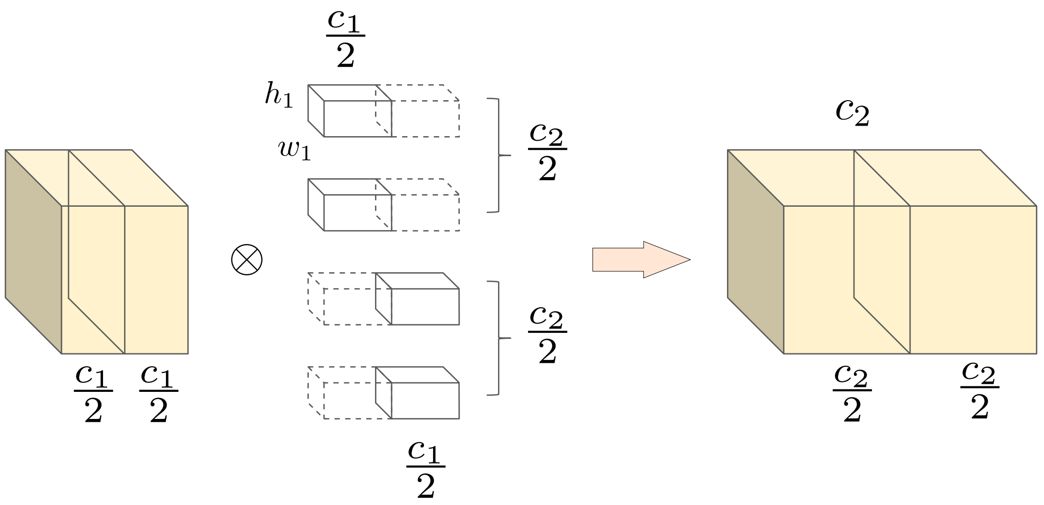


Figure 5.5: Overview of a 2-group convolution operator. Given c_1 feature maps in the input, c_2 filters of size $\frac{c_1}{2} \times h_1 \times w_1$, each output feature map location is the dot product between the filter and corresponding spatial input in the same group.

5.3.2 Circular Convolution

Performing convolutions in the log-polar coordinates is not obvious due to the irregularity of the spatial grid. Notice that the main difference with respect to normal convolution is the periodicity of the signal in the angular dimension but non-periodic in the radial direction. The operation can be implemented by a circular padding before applying conventional convolution operation. The illustration of such operation is shown in Figure 5.6.

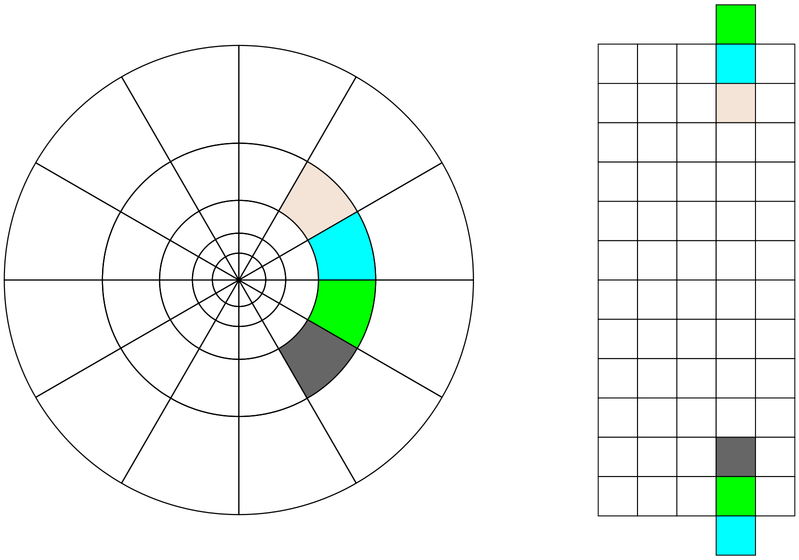


Figure 5.6: The log-polar transform places adjacent bins in spatial domain to two sides of the angular coordinate. Thus convolution operators must take care of this circular nature of the input in the angular coordinates. The blue and green pixels show how circular padding is done prior to a convolution with kernel size 3.

5.4 Network Architecture

For the recognition task we follow the design of state-of-the-art deep networks [52]. The input image is first convolved with a set of 64 filters. Since our foveated input is very small compared to standard networks, we only use kernels size 2x2 instead of 7x7 in this convolutions. After this, the network

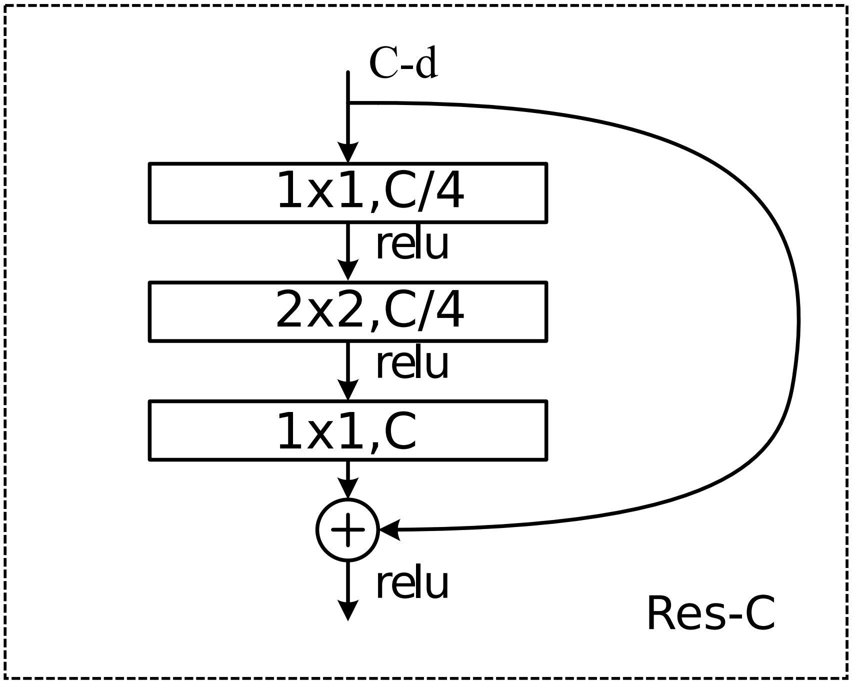


Figure 5.7: Overview of a residual block [52] with number of input channels C. The block includes three convolutions and a residual connection. The last activation function is applied after adding the result with residual connection.

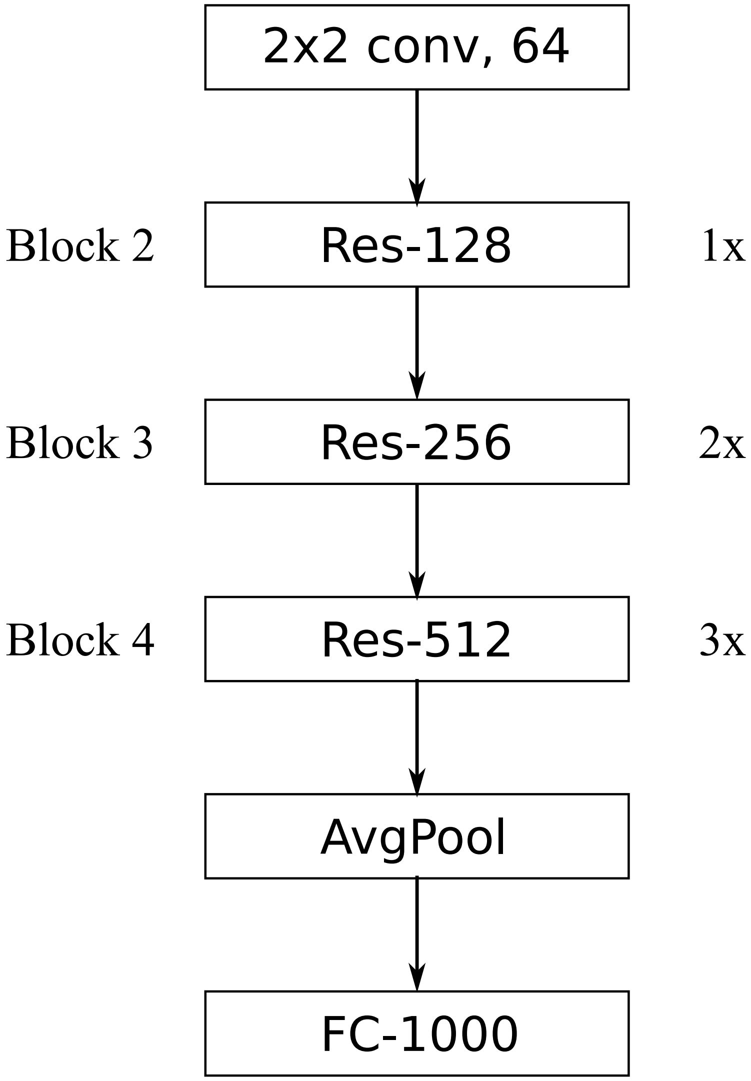


Figure 5.8: Network architecture (a 21-layer Resnet). The network starts with a normal convolution then uses repeated residual blocks (1xBlock2, 2xBlock3 and 3xBlock4). Number of filters doubles every time we reduce (half) the spatial resolution.

replicates many residual blocks. Each residual block consists of a 3 convolutions and a residual connection as shown in Figure 5.7. Each convolution is followed by batch normalization and a rectified non-linearity unit. Spatial dimension reduction is performed by inserting stride-2 convolutions instead of max-pooling. Number of filters is doubled every time we reduce spatial dimension. An average pooling layer is used prior to fully connected layer for classification. We limit the depth of our network to 21 layers due to memory limitation during training. An overview of the network is shown in Figure 5.8.

To accommodate multiple fixations, we extract features from the last convolutional layer and average them across fixations before classification. This process is summarized in Figure 5.9.

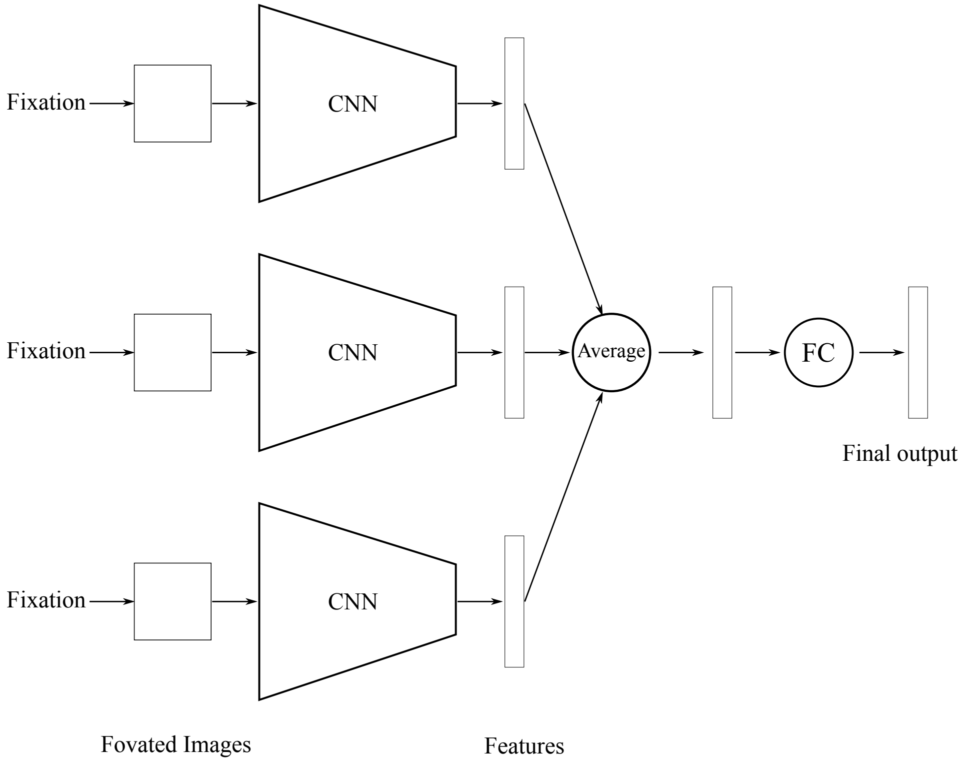


Figure 5.9: For multiple fixations, features from last convolution layers are averaged across fixations before the classification layer.

5.4.1 Training Objective

For each image, the network outputs an activation feature vector $\phi = (\phi_1, \phi_2, \cdots, \phi_C)$ with the same dimension as the number of classes C. This vector is then normalize using softmax fucntion to obtain the likelihood of the input belonging to each class. The probability of the input belonging to class k is given by:

$$p_k = \frac{\exp(\phi_k)}{\sum\limits_{j=1}^{C} \exp(\phi_j)} \tag{5.1}$$

During training the true class label for the input t is also given. In this case, we represent t as a one-hot vector of dimension C (number of classes). The cross entropy loss is then applied to the prediction and corresponding target t:

$$L = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{C} -t_k^{(i)} \log(p_k^{(i)}) \tag{5.2}$$

In the above equation B is the number of samples used in each training iteration (batch size), the superscript i indicates the index of a sample in the batch. To minimize the loss, its gradients with respect to parameters of the networks are computed using backpropagation and stochastic gradient descent is used to update the weights.

For each image we randomly sample fixations near the image center. Then a log-polar transformed input for that fixation is created and is used as input for the network. We use learning rate 10^{-2} and weight decay 10^{-5}. The network is trained for 90 epochs. The learning rate is scheduled to decay by a factor of 10 at the 30-th and 60-th epochs.

5.4.2 Inference

The proposed model uses multiple fixations. In the following experiments we consider two scenarios, one with random fixations and the other with greedy fixations, as explained below.

Random Fixations: We choose the first fixation at the center of the image and subsequent fixations as random around the center of the image.

Greedy Fixations: Gradient based class activation map (GradCam)[102] is a way to evaluate the locations of important features with respect to a specific object class. The overview of this process is outlined in Figure 5.10. First the gradients of the class score with respect to the feature map is computed. Spatial average of the corresponding gradients gives us the weights indicating how importance the features are to that class. The saliency map is then the weighted sum of the feature maps. Given an initial prediction, we generate the top 4 predicted classes based on the class probability scores (equation 5.1) and these class labels are used to generate four corresponding GradCams. The highest score location in each GradCam map is used as new fixations. In total we have 5 fixations to make predictions. An overview of the process is shown in Figure 5.11.

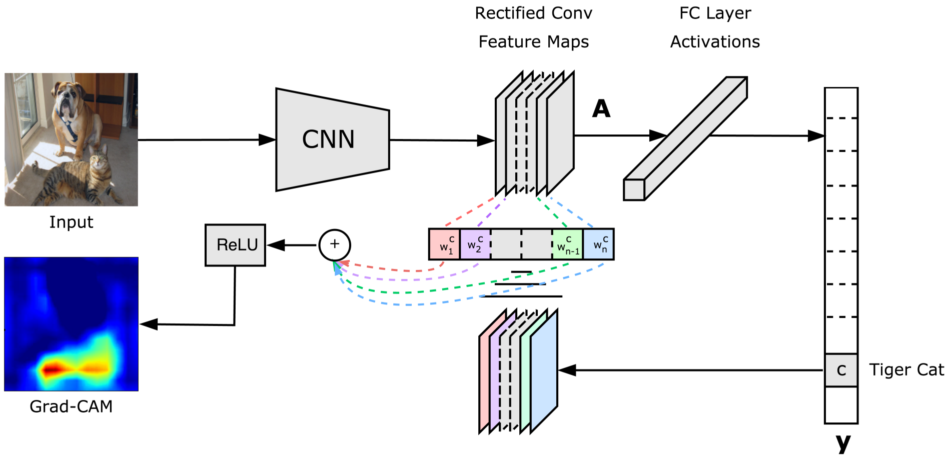


Figure 5.10: Overview of gradient based class activation map (GradCAM) [102]. For a given class and feature map A, the global average pooling of gradients of the class prediction score with respect to the map A is used to weight the importance of individual maps with respect to that class. The weighted sum of activation in the feature map followed by rectified linear activation to obtain Grad-CAM map for the input image for that particualr class.

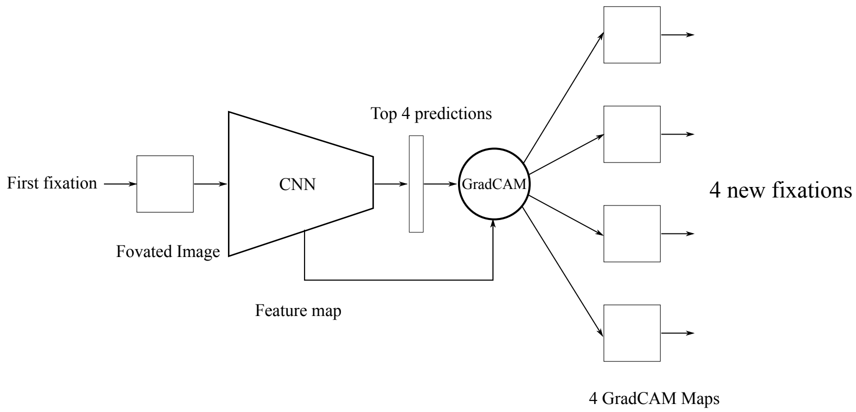


Figure 5.11: Generate fixations in a greedy manner. From the feature map generated from the first fixation. Top4 class prediction are used to create saliency map with GradCAM. For each map the location with maximum score are used as new fixation. In total we have 5 fixations.

5.5 Experiments

We evaluate the results using ImageNet [103] 2012 using top-1 accuracy on the provided validation set. The dataset includes 1.2 million images of 1000 object classes. Current state of the art model uses 152 layers and operates on 224x224 input resolution, achieving 79.8 percent top-1 accuracy [104]. Since we are limited by the existing hardware resources, we use 21 layers in our implementation. In order to make fair comparisons, the baseline models and our proposed architecture all use the same number of layers, with the proposed method using circular convolutions instead of normal convolutions. We use two baselines 1) *1× pixel CNN* using downsampled inputs with the same resolution as foveated images, and 2) *4× pixel CNN* using downsampled inputs with four times the number of pixels as foveated images. The results are reported in Figure 5.12 show that the proposed method at the lowest resolution (single fixation) perform on par with the state-of-the-art models with the same number of down-sampled pixels. With additional fixations, the proposed method outperform normal convolutional models when both models are processing the same number of pixels. Given that in our existing implementation fixations are not optimally generated and integrated, we would expect considerable room for improvement in future research.

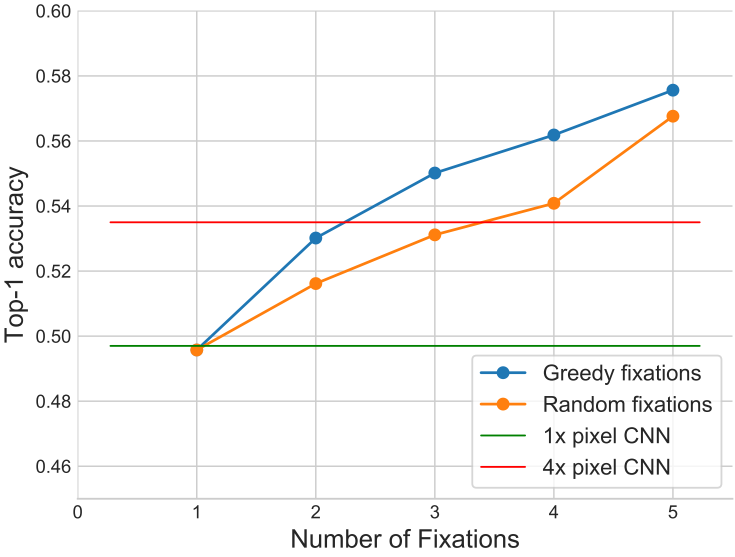


Figure 5.12: Model performance on validation set compared to downsampling baselines (denoted by the horizontal bars). Top-1 accuracy increases with the number of fixations. When using random fixations, model performance at 4 fixations and 1 fixation are similar to downsampling base lines with the same number of pixels. The greedy fixations brings consistent improvements when we have multiple fixations.

5.6 Discussion

We proposed a convolutional network architecture that combines foveated image samples and multiple fixations for object recognition. The primary motivation for this work comes from biological vision systems that effectively combine non-uniform, foveated sampling on the retina, together with eye-movements that scan the scene and sample data at different locations, for scene understanding and recognition. Our results demonstrate the feasibility of foveated image recognition with the potential to outperform current state-of-the-art models. The main contribution is a convolutional neural network that works directly with the foveated input images that achieves competitive recognition rates compared to standard neural networks operating on the same number of input pixels. We also propose an adaptive mechanism trading computation for accuracy without changing the model. Future research directions include utilizing the sequential nature of the fixations for more efficient and effective recognition models.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The thesis investigated the requirements and implementations that could support active foveated vision. Towards this we carried out extensive human subject experiments to understand how humans perform visual search. Results from the experiments called for modeling human eye tracking data directly. We proposed neural networks models for gaze prediction that outperform previous works. Lastly, we proposed a convolutional network that could directly operate on foveated input images for object recognition. The main contributions of the thesis are summarized below:

- A comprehensive study of how humans search for target people in dynamic scenes (chapter 3) is presented. A novel video database is created to test how different parts of a person contribute towards eye-movements and person identification. We find subjects rely strongly on the face where performance drops by a large amount when facial information is removed. Another finding is that the search errors can dominate the overall recognition accuracy in human subject experiments. The recorded eye movements show that humans have strong a bias towards faces. However, when forced to fixate on faces, subjects do not obtain maximum performance, which suggests that face-centered strategy might not necessarily maximize the person identification performance in human subject trials but likely arises as a byproduct of the implementation of a heuristic strategy that optimizes perceptual performance across a battery of evolutionary important tasks. Performance of two current computer models, a foveated ideal observer and a naive convolutional neural network, is compared against human, showing that machine models treat faces similar to other features and are outperformed by human subjects by a large margin.

- We propose a neural model to predict human eye movement, trained directly on human subjects data (chapter 4). Given an image, the model predicts the most likely sequence of fixations a human would follow. In developing this model we leverage recent advances in image recognition using convolutional neural networks and sequence modeling with recurrent neural networks. Feature maps from convolutional neural networks are used as inputs to a recurrent neural network. The recurrent neural network acts like a visual working memory that integrates the scene information and outputs a sequence of fixations. The model is trained on human eye tracking data. The proposed approach removes dataset-dependent feature engineering steps and achieves state of the art performance.

- Finally, we develop an image recognition model that operates on foveated input images (chapter 5). This is in contrast with current methods that use spatially variant filtering to create foveated images, retaining the same number of pixels as the original inputs. Assuming a log-polar representation of foveated sensory signals, a circular convolutional neural network is designed to perform image recognition. To the best of our knowledge, this is the first time that a convolutional network is developed to work directly with the foveated image dat. The proposed method is also able to handle multiple fixations giving better performance with more fixations, thus adaptive to the given computational budget.

6.2 Future Work

6.2.1 Foveated Image Recognition with Gazing

The main limitation of our fixation prediction model is that it is not operating on foveated input. At the same time, our foveated recognition model is not taking sequential nature of fixations. A combination of a foveated image recognition with fixation planing and integration would be ideal. This can be done using a recurrent model integrate information from each fixation and output both the next location and class probabilities at the same time.

6.2.2 Foveated Object Detection

A natural extension of the work presented in Chapter 5 is to perform object detection with foveated images. Object detection is an expensive visual search task, requiring multiple stages of computation [105]. The model proposed in the previous section could be further extended to output the spatial extent of the object at current fixation, a boundind box for example. However we need a mechanism to determine when an object is detected or more exploration is necessary. This could be implemented similarly to what currently done in machine translation where the model predict a STOP token indicating the end of a sentence.

6.2.3 Learned Foveated Filtering

Most current work on foveated filtering makes assumptions about the nature of the spatially varying filters based on either biological experiments (for example, the FIO models described in Chapter 3, or computational models such as those based on Gabor filtering.) An alternate approach would be to learn these filters from data just like any convolutional filters in deep learning architectures. One can approximate the process by sampling with high resolution log-polar space and then perform spatially varying filtering. The learned kernels represent the optimal foveated filtering for a specific task. This can be useful for human vision research to understand how low-level visual processing is done and for a better understanding of the further understanding the notion of metamers introduced in [95]. Figure 6.1 shows a set of spatially varying filters trained using imagenet dataset to maximize recognition accuracy, and it would be interesting to explore this further in the context of foveated recognition described in Chapter 5.

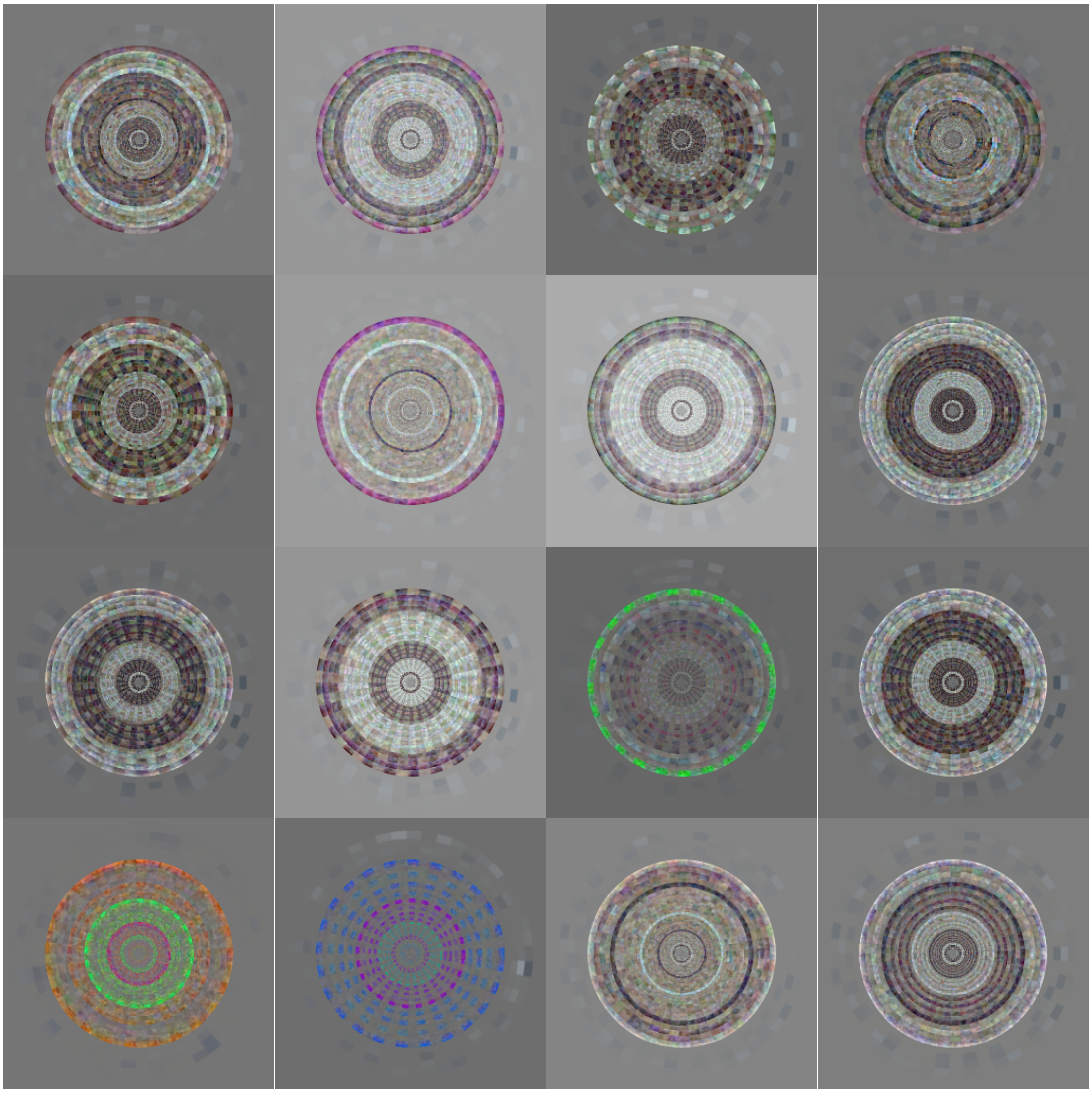


Figure 6.1: Visualization of learned (spatially varying) filters

Bibliography

[1] J. L. Crowley, P. Bobet, and M. Mesrabi, *Gaze control for a binocular camera head*, in *European Conference on Computer Vision*, pp. 588–596, Springer, 1992.

[2] A. S. Rojer and E. L. Schwartz, *Design considerations for a space-variant visual sensor with complex-logarithmic geometry*, in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. 2, pp. 278–285, IEEE, 1990.

[3] A. Califano, R. Kjeldsen, and R. M. Bolle, *Data and model driven foveation*, in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. 1, pp. 1–7, IEEE, 1990.

[4] B. Scassellati, *A binocular, foveated active vision system*, tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1999.

[5] K. Pahlavan, T. Uhlin, and J.-O. Eklundh, *Integrating primary ocular processes*, in *European Conference on Computer Vision*, pp. 526–541, Springer, 1992.

[6] R. W. Rodieck and R. W. Rodieck, *The first steps in seeing*, vol. 1. Sinauer Associates Sunderland, MA, 1998.

[7] L. Itti, C. Koch, and E. Niebur, *A model of saliency-based visual attention for rapid scene analysis*, *IEEE Transactions on Pattern Analysis & Machine Intelligence* (1998), no. 11 1254–1259.

[8] J. Harel, C. Koch, and P. Perona, *Graph-based visual saliency*, in *Advances in neural information processing systems*, pp. 545–552, 2006.

[9] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, *A coherent computational approach to model bottom-up visual attention*, *Pattern*

Analysis and Machine Intelligence, IEEE Transactions on **28** (2006), no. 5 802–817.

[10] M. P. Eckstein, *Probabilistic computations for attention, eye movements, and search*, *Annual review of vision science* **3** (2017) 319–342.

[11] J. M. Findlay and V. Brown, *Eye scanning of multi-element displays: I. scanpath planning*, *Vision research* **46** (2006), no. 1-2 179–195.

[12] G. J. Zelinsky, *A theory of eye movements during target acquisition.*, *Psychological review* **115** (2008), no. 4 787.

[13] J. Najemnik and W. S. Geisler, *Eye movement statistics in humans are consistent with an optimal search strategy*, *Journal of Vision* **8** (2008), no. 3 4–4.

[14] Wikipedia, *Visual search*, *URL: https://en.wikipedia.org/wiki/* (2015) 20.

[15] K. Fukushima and S. Miyake, *Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition*, in *Competition and cooperation in neural nets*, pp. 267–285. Springer, 1982.

[16] D. H. Hubel and T. N. Wiesel, *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex*, *The Journal of physiology* **160** (1962), no. 1 106–154.

[17] E. T. Rolls, *Invariant visual object and face recognition: neural and computational bases, and a model, visnet*, *Frontiers in Computational Neuroscience* **6** (2012) 35.

[18] M. Riesenhuber and T. Poggio, *Neural mechanisms of object recognition*, *Current opinion in neurobiology* **12** (2002), no. 2 162–168.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE* **86** (1998), no. 11 2278–2324.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, *Backpropagation applied to handwritten zip code recognition*, *Neural computation* **1** (1989), no. 4 541–551.

[21] K. Kuniyoshi, N. Kita, K. Sugimoto, S. Nakamura, and T. Suehiro, *A foveated wide angle lens for active vision*, in *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, vol. 3, pp. 2982–2988, IEEE, 1995.

[22] D. G. Bailey and C.-S. Bouganis, *Reconfigurable foveated active vision system*, in *Sensing Technology, 2008. ICST 2008. 3rd International Conference on*, pp. 162–167, IEEE, 2008.

[23] D. Young, *Straight lines and circles in the log-polar image.* School of Cognitive and Computing Sciences, University of Sussex, 2000.

[24] C. A. Hahn, P. J. Phillips, and A. J. O'Toole, *Contributing factors of person recognition in natural environments*, *Journal of Vision* **14** (2014), no. 10 560–560.

[25] A. Rice, P. J. Phillips, V. Natu, X. An, and A. J. OToole, *Unaware person recognition from the body when face identification fails*, *Psychological Science* **24** (2013), no. 11 2235–2243.

[26] A. Rice, P. J. Phillips, and A. O'Toole, *The role of the face and body in unfamiliar person identification*, *Applied Cognitive Psychology* **27** (2013), no. 6 761–768.

[27] N. Simhi and G. Yovel, *The contribution of the body and motion to whole person recognition*, *Vision research* **122** (2016) 12–20.

[28] J. M. Wolfe and T. S. Horowitz, *Five factors that guide attention in visual search*, *Nature Human Behaviour* **1** (2017), no. 3 0058.

[29] M. P. Eckstein, *Visual search: A retrospective*, *Journal of vision* **11** (2011), no. 5 14–14.

[30] W. Estes and H. Taylor, *A detection method and probabilistic models for assessing information processing from brief visual displays*, *Proceedings of the National Academy of Sciences* **52** (1964), no. 2 446–454.

[31] R. M. Shiffrin and W. Schneider, *Automatic and controlled processing revisited.*, .

[32] A. Treisman, *Search, similarity, and integration of features between and within dimensions.*, *Journal of Experimental Psychology: Human Perception and Performance* **17** (1991), no. 3 652.

[33] M. M. Chun and Y. Jiang, *Top-down attentional guidance based on implicit learning of visual covariation*, *Psychological Science* **10** (1999), no. 4 360–365.

[34] J. Najemnik and W. S. Geisler, *Optimal eye movement strategies in visual search*, *Nature* **434** (2005), no. 7031 387.

[35] M. P. Eckstein, W. Schoonveld, S. Zhang, S. C. Mack, and E. Akbas, *Optimal and human eye movements to clustered low value cues to increase decision rewards during search*, *Vision research* **113** (2015) 137–154.

[36] A. E. Burgess and H. Ghandeharian, *Visual signal detection. ii. signal-location identification*, *JOSA A* **1** (1984), no. 8 906–910.

[37] F. O. Bochud, C. K. Abbey, and M. P. Eckstein, *Search for lesions in mammograms: statistical characterization of observer responses*, *Medical Physics* **31** (2004), no. 1 24–36.

[38] M. Michel and W. S. Geisler, *Intrinsic position uncertainty explains detection and localization performance in peripheral vision*, *Journal of Vision* **11** (2011), no. 1 18–18.

[39] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search.*, *Psychological review* **113** (2006), no. 4 766.

[40] M. J. Bravo and H. Farid, *The specificity of the search template*, *Journal of Vision* **9** (2009), no. 1 34–34.

[41] K. Koehler and M. P. Eckstein, *Beyond scene gist: Objects guide search more than scene background.*, *Journal of Experimental Psychology: Human Perception and Performance* **43** (2017), no. 6 1177.

[42] J. M. Wolfe and T. S. Horowitz, *What attributes guide the deployment of visual attention and how do they do it?*, *Nature reviews neuroscience* **5** (2004), no. 6 495.

[43] M. F. Peterson and M. P. Eckstein, *Looking just below the eyes is optimal across face recognition tasks*, *Proceedings of the National Academy of Sciences* **109** (2012), no. 48 E3314–E3323.

[44] J. M. Henderson, C. C. Williams, and R. J. Falk, *Eye movements are functional during face learning*, *Memory & cognition* **33** (2005), no. 1 98–106.

[45] J. Arizpe, D. J. Kravitz, G. Yovel, and C. I. Baker, *Start position strongly influences fixation patterns during face processing: Difficulties with eye movements as a measure of information use*, *PloS one* **7** (2012), no. 2 e31106.

[46] M. F. Peterson and M. P. Eckstein, *Learning optimal eye movements to unusual faces*, *Vision research* **99** (2014) 57–68.

[47] G. Van Belle, M. Ramon, P. Lefèvre, and B. Rossion, *Fixation patterns during recognition of personally familiar and unfamiliar faces*, *Frontiers in psychology* **1** (2010) 20.

[48] P. G. Schyns, L. Bonnar, and F. Gosselin, *Show me the features! understanding recognition from the use of visual information*, *Psychological science* **13** (2002), no. 5 402–409.

[49] C. Vinette, F. Gosselin, and P. G. Schyns, *Spatio-temporal dynamics of face recognition in a flash: It's in the eyes*, *Cognitive Science* **28** (2004), no. 2 289–301.

[50] J. Yuen, B. Russell, C. Liu, and A. Torralba, *Labelme video: Building a video database with human annotations*, in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1451–1458, IEEE, 2009.

[51] Y. Tsank and M. P. Eckstein, *Domain specificity of oculomotor learning after changes in sensory processing*, *Journal of Neuroscience* (2017) 1208–17.

[52] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[53] C. Cortes and V. Vapnik, *Support-vector networks*, *Machine learning* **20** (1995), no. 3 273–297.

[54] B. A. Wandell, *Foundations of vision.* Sinauer Associates, 1995.

[55] R. Desimone and J. Duncan, *Neural mechanisms of selective visual attention*, *Annual review of neuroscience* **18** (1995), no. 1 193–222.

[56] S. K. Ungerleider and L. G, *Mechanisms of visual attention in the human cortex*, *Annual review of neuroscience* **23** (2000), no. 1 315–341.

[57] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part-based models*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **32** (2010), no. 9 1627–1645.

[58] D. Walther and C. Koch, *Modeling attention to salient proto-objects*, *Neural networks* **19** (2006), no. 9 1395–1407.

[59] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, *Simulating human saccadic scanpaths on natural images*, in *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pp. 441–448, IEEE, 2011.

[60] H. Liu, D. Xu, Q. Huang, W. Li, M. Xu, and S. Lin, *Semantically-based human scanpath estimation with hmms*, in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 3232–3239, Dec, 2013.

[61] M. Jiang, X. Boix, J. X. Gemma Roig, L. V. Gool, and Q. Zhao, *Learning to predict sequences of human visual fixations*, *Neural Networks, IEEE Transactions on* (2015).

[62] A. Borji, H. Tavakoli, D. Sihite, and L. Itti, *Analysis of scores, datasets, and models in visual saliency prediction*, in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 921–928, Dec, 2013.

[63] N. Bruce and J. Tsotsos, *Saliency based on information maximization*, in *Advances in neural information processing systems*, pp. 155–162, 2005.

[64] 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** (2008), no. 7 32.

[65] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, *Visual saliency based on conditional entropy*, in *Computer Vision–ACCV 2009*, pp. 246–257. Springer, 2010.

[66] X. Hou and L. Zhang, *Dynamic visual attention: Searching for coding length increments*, in *Advances in neural information processing systems*, pp. 681–688, 2009.

[67] R. J. Peters and L. Itti, *Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention*, in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, 2007.

[68] J. Li, Y. Tian, T. Huang, and W. Gao, *Probabilistic multi-task learning for visual saliency estimation in video*, *International journal of computer vision* **90** (2010), no. 2 150–165.

[69] A. Borji, D. N. Sihite, and L. Itti, *Probabilistic learning of task-specific visual attention*, in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 470–477, IEEE, 2012.

[70] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, *Learning to detect a salient object*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **33** (2011), no. 2 353–367.

[71] J. Yang and M.-H. Yang, *Top-down visual saliency via joint crf and dictionary learning*, in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 2296–2303, IEEE, 2012.

[72] T. Judd, K. Ehinger, F. Durand, and A. Torralba, *Learning to predict where humans look*, in *Computer Vision, 2009 IEEE 12th international conference on*, pp. 2106–2113, IEEE, 2009.

[73] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, *Predicting human gaze beyond pixels*, *Journal of vision* **14** (2014), no. 1 28.

[74] R. Zhao, W. Ouyang, H. Li, and X. Wang, *Saliency detection by multi-context deep learning*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265–1274, 2015.

[75] X. B. Xun Huang, Chengyao Shen and Q. Zhao, *Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks*, in *Computer Vision, 2015 IEEE 18th international conference on*, IEEE, 2015.

[76] D. Brockmann and T. Geisel, *The ecology of gaze shifts*, *Neurocomputing* **32** (2000) 643–650.

[77] G. Boccignone and M. Ferraro, *Modelling gaze shift as a constrained random walk*, *Physica A: Statistical Mechanics and its Applications* **331** (2004), no. 1 207–218.

[78] I. Sutskever, O. Vinyals, and Q. V. V. Le, *Sequence to sequence learning with neural networks*, in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), pp. 3104–3112. Curran Associates, Inc., 2014.

[79] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, *Show and tell: A neural image caption generator*, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 3156–3164, June, 2015.

[80] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, *Long-term recurrent convolutional networks for visual recognition and description*, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 2625–2634, June, 2015.

[81] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 3431–3440, June, 2015.

[82] N. Sunderhauf, C. McCool, B. Upcroft, and P. Tristan, *Fine-grained plant classification using convolutional neural networks for feature extraction*, in *Working notes of CLEF 2014 conference*, 2014.

[83] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, *Efficient object localization using convolutional networks*, in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 648–656, June, 2015.

[84] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[85] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, *Neural computation* **9** (1997), no. 8 1735–1780.

[86] P. M. Bays and M. Husain, *Active inhibition and memory promote exploration and search of natural scenes*, *Journal of Vision* **12** (2012), no. 8 8.

[87] T. Sauer, J. A. Yorke, and M. Casdagli, *Embedology*, *Journal of statistical Physics* **65** (1991), no. 3-4 579–616.

[88] H. Jarodzka, K. Holmqvist, and M. Nyström, *A vector-based, multidimensional scanpath similarity measure*, in *Proceedings of the 2010 symposium on eye-tracking research & applications*, pp. 211–218, ACM, 2010.

[89] L. Gupta, D. L. Molfese, R. Tammana, and P. G. Simos, *Nonlinear alignment and averaging for estimating the evoked potential*, *Biomedical Engineering, IEEE Transactions on* **43** (1996), no. 4 348–356.

[90] C. M. Privitera and L. W. Stark, *Algorithms for defining visual regions-of-interest: Comparison with eye fixations*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22** (2000), no. 9 970–982.

[91] S. B. Needleman and C. D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, *Journal of molecular biology* **48** (1970), no. 3 443–453.

[92] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: A simple way to prevent neural networks from overfitting*, *The Journal of Machine Learning Research* **15** (2014), no. 1 1929–1958.

[93] T. Tieleman and G. Hinton, *Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude*, in *Coursera: Neural Networks for Machine Learning*, 2012.

[94] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, *Theano: a CPU and GPU math expression compiler*, in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June, 2010. Oral Presentation.

[95] J. Freeman and E. P. Simoncelli, *Metamers of the ventral stream*, *Nature neuroscience* **14** (2011), no. 9 1195–1201.

[96] W. S. Geisler and J. S. Perry, *Real-time foveated multiresolution system for low-bandwidth video communication*, in *Human vision and electronic imaging III*, vol. 3299, pp. 294–306, International Society for Optics and Photonics, 1998.

[97] C. Melıcio, R. Figueiredo, A. F. Almeida, A. Bernardino, and J. Santos-Victor, *Object detection and localization with artificial foveal visual attention*, .

[98] E. Akbas and M. P. Eckstein, *Object detection through search with a foveated visual system*, *PLOS Computational Biology* **13** (2017), no. 10 e1005743.

[99] P. Kortum and W. S. Geisler, *Implementation of a foveated image coding system for image bandwidth reduction*, in *Human Vision and Electronic Imaging*, vol. 2657, pp. 350–361, International Society for Optics and Photonics, 1996.

[100] H. Yamamoto, Y. Yeshurun, and M. D. Levine, *An active foveated vision system: Attentional mechanisms and scan path covergence measures*, *Computer Vision and Image Understanding* **63** (1996), no. 1 50–65.

[101] H. P. Moravec, *Techniques towards automatic visual obstacle avoidance*, .

[102] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, *Grad-cam: Visual explanations from deep networks via gradient-based localization.*, in *ICCV*, pp. 618–626, 2017.

[103] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, in *CVPR09*, 2009.

[104] S. Xie, R. Girshick, P. Dollr, Z. Tu, and K. He, *Aggregated residual transformations for deep neural networks*, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017*, 2016.

[105] K. He, G. Gkioxari, P. Dollár, and R. Girshick, *Mask r-cnn*, in *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988, IEEE, 2017.