Understanding and predicting where people look in images

by

Tilke Judd

B.S., Massachusetts Institute of Technology (2003)
S.M., Massachusetts Institute of Technology (2007)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011

Author .....

Department of
Electrical Engineering and Computer Science
May 20, 2011

Certified by .....
Frédo Durand
Associate Professor
Thesis Supervisor

Certified by... .....
Antonio Torralba
Associate Professor
Thesis Supervisor

Accepted by .....
Professor Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Students

Understanding and predicting where people look in images

by

Tilke Judd

Submitted to the Department of
Electrical Engineering and Computer Science
on May 20, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

For many applications in graphics, design, and human computer interaction, it is essential to understand where humans look in a scene. This is a challenging task given that no one fully understands how the human visual system works. This thesis explores the way people look at different types of images and provides methods of predicting where they look in new scenes. We describe a new way to model where people look from ground truth eye tracking data using techniques of machine learning that outperforms all existing models, and provide a benchmark data set to quantitatively compare existing and future models. In addition we explore how image resolution affects where people look. Our experiments, models, and large eye tracking data sets should help future researchers better understand and predict where people look in order to create more powerful computational vision systems.

Thesis Supervisor: Frédo Durand
Title: Associate Professor

Thesis Supervisor: Antonio Torralba
Title: Associate Professor

Acknowledgments

I like to say that I came to MIT for undergraduate studies because MIT had the best ballroom dance team. MIT does have a top ballroom team, and I enjoyed dancing with them for several years. But my experience at MIT has gone far beyond that – MIT has provided me immense opportunities to learn and grow. As an undergrad, I studied math, taught in China, learned to sail. I participated in the UROP program which gave me my first exposure to research. I met Professor Muriel Medard who recommended graduate school, Frédo and the MIT graphics group. In graduate school, opportunities continued – opportunities to write papers, to give talks, to travel to conferences, to become a researcher. There were so many great people that helped along the way. I would like to acknowledge and to send thanks to a few of them:

- To Frédo Durand – for guiding me as a researcher and as a person. You have always had my best interests in mind when providing advice. I thank you for the 12 Siggraph practice talks you coached me through, for the feedback you provided on papers in record time, and for the way you pause to let me write thoughts down in meetings. Thank you for creating a group with a positive, productive ambiance. One's experience in graduate school is so highly affected by one's advisor. My experience was great.

- To Antonio Torralba – for showing me how to think about research problems. You have such a genuine interest and curiosity in the problems you work on and I admire the way you tinker with them until they unfold. Not only do you aim to understand what is going on, you also find ways to display it in visually compelling and informative ways. Your enthusiasm for research is contagious.

- To Aude Oliva – for welcoming me into the brain and cognitive science community and introducing me to human subject eye tracking tests. You removed so many obstacles, allowing me to be productive. I really enjoyed collaborating with you and your group.

- To Bill Freeman - for your simple advice to "do good work".

- To colleagues of mine who's help directly affected my thesis work: Barbara Hidalgo-Sotelo for help with eye tracking, Krista Ehinger for her statistics prowess, Nicolas Pinto and Yann Le Tallec for advice on machine learning.

- To the MIT Graphics Group. I was blessed to arrive in a group that has such an engaging and positive environment. As graduate students together we exchanged ideas on research, 4chan, portal, climbing, LaTeX, squash, diabetes, bidirectional texture functions, glassblowing, dynamic programming, shoes, technical interviews, salmiakki, MATLAB commands, just to name a few.

- To Super Bryt Bradley – for making everything work. The world runs because Bryt keeps it running.

- To my formal or informal mentors during graduate school: Muriel Medard, Luca Daniel, Ruth Rosenholtz, Cary Philipps from Industrial Light and Magic, Rob Rolleston from Xerox, Joyce Wu. Thanks for your perspective, advice and guidance.

- To my friends from MIT – you are some of the most adventurous, intelligent, curious, driven, and immensely inspiring classmates. I probably learned the most from you. I have great memories of interacting with you through the Campus Movie Festival, Leaders in Life program, and Graduate Women @ MIT conferences.

- To my friends and housemates at 110 Pleasant Street – for keeping me sane and balanced. 110 Pleasant has been my home and family away from home. I will look back very fondly at our time together.

- To my parents – thanks for all the "Why are potholes round?" questions at the dinner table and for so strongly emphasizing education. I understood the priorities right away: no money for excessive clothes, plenty for textbooks or music lessons. Rather than good cars, we all get great university educations.

- To my Daddy who did his PhD while I was a little girl. I remember him waking me and my sisters up at night to show us his finished published PhD. Look Daddy, now I'm a Dr.! I am in his dedication and he is in mine.

- To my sisters – we've always done a great job at motivating each other to excel, and we all keep the bar high! I'm happy we can share accomplishments.

- To my family – for all your love. It has given me the confidence to be who I am, and it keeps us tight though we are spread across continents.

- To Yann – who I met in graduate school. You keep me endlessly happy.

Contents

Chapter 1

Introduction

A red scarf in a sea of black tuxedos catches your attention. You notice faces when looking for a friend in a crowd. Bikers are more likely to notice bike lanes. In all these situations, the mechanism in the brain that determines what part of the incoming sensory data is currently of most interest is called *selective attention.* A definition was first outlined by James [1890] and here defined by [Corbetta 1990]: "Attention defines the mental ability to select stimuli, responses, memories or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant." Selective attention exists for all senses and has developed through evolution because of the human need to deal with an overwhelming amount of sensory input at each moment. The amount of data is too high to be fully processed in details; the brain must prioritize.

Visual attention is often compared to a spotlight in a dark room. The fovea–the center of the retina–has the highest resolution. Areas around the fovea fall off in resolution. Directing the gaze to a certain region corresponds to directing a spotlight to a certain part of the room [Shulman et al. 1979]. By moving the spotlight around, a person can get an impression of the room, and similarily, by scanning the scene with quick eye movements, one gets a detailed impression of it. We move our fovea to specific areas of interest to fixate on them. We then move, or saccade, to the next location of interest.

People were only able to measure these fixations and saccades with the invention of eye trackers. The first non-intrusive eye trackers were built by Guy Thomas Buswell in Chicago in the 1920s and used beams of light that were reflected on the eye and then recorded on film [Buswell, 1922]. In the 1950s, Alfred L. Yarbus [1967] performed important eye tracking research and wrote about the critical relation between fixations and interest:

> "Records of eye movements show that the observer's attention is usually held only by certain elements of the picture.... Eye movement reflects the human thought processes; so the observer's thought may be followed to some extent from records of eye movement (the thought accompanying the examination of the particular object). It is easy to determine from these records which elements attract the observer's eye (and, consequently, his

thought), in what order, and how often."

Following this vein, Just and Carpenter [1980] formulated the influential *Strong Eye-Mind Hypothesis*, the hypothesis that "there is no appreciable lag between what is fixated and what is processed". If this hypothesis is correct, then when a subject looks at a word or object, he or she also thinks about it for as long as the recorded fixation. Subsequent research questioned this hypothesis in light of *covert attention*, or the ability to pay attention to something one is not looking at. According to Hoffman [1998], current consensus is that visual attention is always slightly (100 to 250 ms) ahead of the eye. Despite this, the eye-mind hypothesis is still a useful and commonly made assumption in eye tracking and visual attention research.

Yarbus also showed that the task given to a subject has a very large influence on the subject's eye movement. He had observers view Repin's painting "The Unexpected Visitor" under several different task conditions and saw that scan paths differed greatly (see Fig. 1-1).

> The cyclical pattern in the examination of pictures "is dependent not only on what is shown on the picture, but also on the problem facing the observer and the information that he hopes to gain from the picture."

This behavior is due to the interaction of two major mechanisms of visual attention: *bottom-up factors* and *top-down factors*. Bottom-up factors are derived solely from the visual scene [Nothdurft 2005]. Regions that attract our attention are called *salient* and are sufficiently different with respect to their surrounding. This attentional mechanism is also called *automatic, reflexive, or stimulus-driven*. The second mechanism, top-down attention, is driven by cognitive factors such as knowledge, expectations and current goals [Corbetta and Shulman 2002]. This is also called *goal-driven, voluntary*, or *centrally cued* attention. During normal human perception, both mechanisms interact. As per Theeuwes [2004], bottom-up influence is not easy to suppress: a highly salient region captures the focus of your attention regardless of the task.

At the same time that our understanding of the human visual system has increased dramatically, many other fields of research have become very interested in this natural process of selective attention. Computer vision systems have to deal with sometimes millions of pixel values from each frame and the computational complexity of many problems related to the interpretation of image data is very high [Tsotsos 1987]. The task becomes especially difficult when the system has to operate in real-time. In order to cope with these requirements, researchers in computer vision, graphics and robotics have investigated how the concepts of human selective attention can be exploited to prioritize information.

Attentional models find regions of interest which help identify segments for image segmentation, drive adaptive levels of image compression, help coordinate points of interest in image matching. Attentional models can help robots with place recognition and localization and are the basis for active vision which help robots decide where to look next. A plethora of graphics applications benefit from understanding

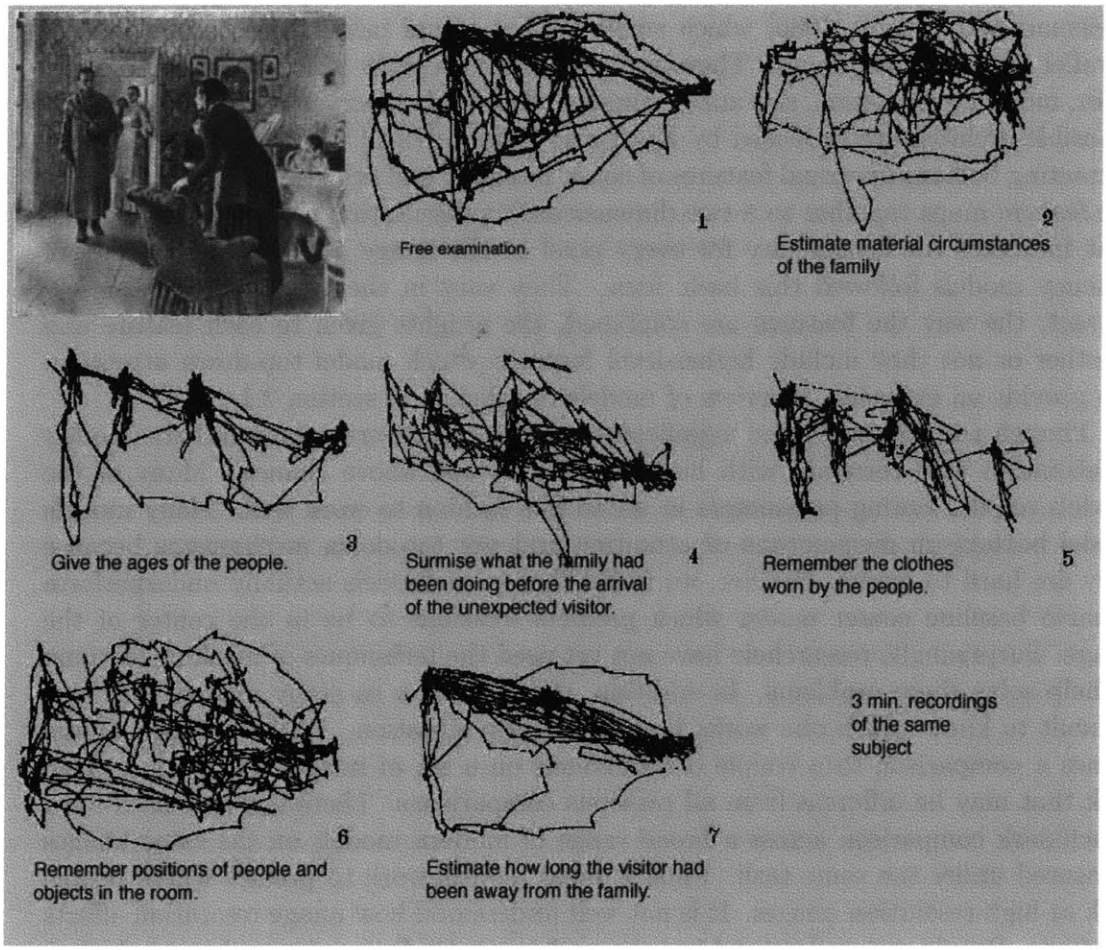

Figure 1-1: Yarbus famously showed that different tasks given to a subject viewing Repin's painting"The Unexpected Visitor" resulted in very different eye movements.

regions of interest in an image: automatic cropping, driving level of detail in abstracted or stylized image, and image retargeting. Understanding where people look helps designers design visual content and market researchers better understand how people interact with their content. We give a more thorough review of applications of selective attention in section 2.2.

Both the need for computational models, and a better understanding of the human visual system led to the development of computational models of visual attention. Many of the models are built around the Feature Integration Theory (FIT) from Treisman and Gelade [1980] which suggests that visual information is analyzed in parallel from different maps. There is a map for each early visual feature including color, motion, orientation, size and luminance. From this theory was born the neurally plausible architecture proposed by Koch and Ullman [1985]. This model consists of extracting bottom-up visual features of color, intensity and orientation and combining the feature maps together to a two-dimensional representation called a saliency map that indicates the conspicuity for every pixel in the image. An extensive series of saliency models followed this basic idea. They vary in the types of features they extract, the way the features are combined, the weights given to each feature and whether or not they include higher-level features which model top-down attention. We provide an extensive overview of models of saliency in section 2.1.

Though research advanced significantly in the last several years and the models of attention now correlate with human fixations well above chance. Many of the models require tuning parameters in an ad hoc fashion to work well. Many models model bottom-up mechanisms of attention and not top-down mechanisms because they are hard to model. Further, we found that some models actually underperform a naïve baseline center model which predicts fixations to be in the center of the image. Surprisingly, researchers have not yet used the techniques of machine learning to help solve these problems. In addition, there are now so many models that it is difficult to know which one works best for a given situation. With each new model comes a comparison to a couple other models on a set of images and with a given task that may be different from all previous comparisons. There is a real need for a benchmark comparison across a broad range of modern models on the same images measured under the same task. Finally, most models work to predict where people look at high resolution images. It is not well understood how image resolution affects fixations. People can understand images, and recognize faces, at very low resolution. In these situations are people looking at locations that give them the most information about the image–and these locations the same across resolutions? For computational efficiency reasons, it is also beneficial to understand how low image resolution can be before observers' fixation patterns are fundamentally altered.

Contributions In this thesis, we help solve these problems. Specifically, we make the following contributions:

1. **We create a new model that predicts where people look in images.** Our model is learned directly from ground-truth data by applying techniques of machine learning and is described in chapter 3. This approach enables us

to learn weights of features for a model automatically. The model incorporates several top-down features and models the human bias to look towards the center. The model outperforms currently available models of saliency, and the naïve center model, in predicting where people look on natural photographs and gets 90 percent of the way to human performance. This work is also published in [Judd *et al.*, 2009].

2. **We provide a benchmark to compare models of saliency.** It is critical to have a quantitative way to evaluate progress. In chapter 4, we describe the benchmark data set and measurement critieria we use to compare ten modern models of saliency. We measure how well each model performs at predicting where people look under three different metrics. We also set up a platform to evaluate future models through an online website[1].

3. **We explore how people look at low resolution images.** In chapter 5 we observe where people look as the resolution of images decreases and analyse specific trends surrounding the bias of fixations toward the center, the number and frequency of fixations, the consistency of fixations across and between different resolutions. We show how fixation consistency is related to image complexity and hypothesize that this is related to image understanding. This work is also published in [Judd *et al.*, 2011].

4. **We provide three new eye tracking data sets.** We make all of the stimuli and fixation data retrieved from our experiments available to the public online.

[1] http://people.csail.mit.edu/tjudd/SaliencyBenchmark

Chapter 2

Background

In this chapter we start by exploring the space of existing computational models of visual attention both by understanding the general structure common to most models and then outlining at groups of models in rough chronological order. We then look at the applications of attentional models in section 2.2, and review the databases of eye tracking data available for evaluating and comparing saliency models in section 2.3. In addition we describe the experimental protocol we used for all the eye tracking experiments in this thesis in section 2.4.

2.1 Computational models of visual attention

In computer vision, robotics, human-computer interaction, computer graphics and design, there is an increasing interest in a mechanism that selects the most relevant parts within a large amount of visual data that is modeled after the human visual system. The systems are all built on the psychological theories of the human visual system, but in the end, they have an engineering objective. The objective in making these models is often both to understand human perception, and to improve vision systems and end up with a way of computing a numerical value of the likelihood of attending to, or the saliency of, every location in an image. Usually these models are able to cope with both synthetic images and natural scenes. The systems vary in detail but have a similar structure. First we describe the similar structure and then group important models by type and describe them in rough chronological order.

2.1.1 General structure of computational models

Most computational models of attention have a similar structure, which is depicted in Figure 2-1. The structure is adapted from Feature Integration Theory [Treisman and Gelade, 1980] and the Guided Search model [Wolfe *et al.*, 1989] and appears first in the algorithmic model of attention by Koch and Ullman [1985]. The main idea is to compute several features in parallel and to fuse their saliencies in a representation which is usually called a *saliency map*.

More specifically, the models generally include the following steps: First, the

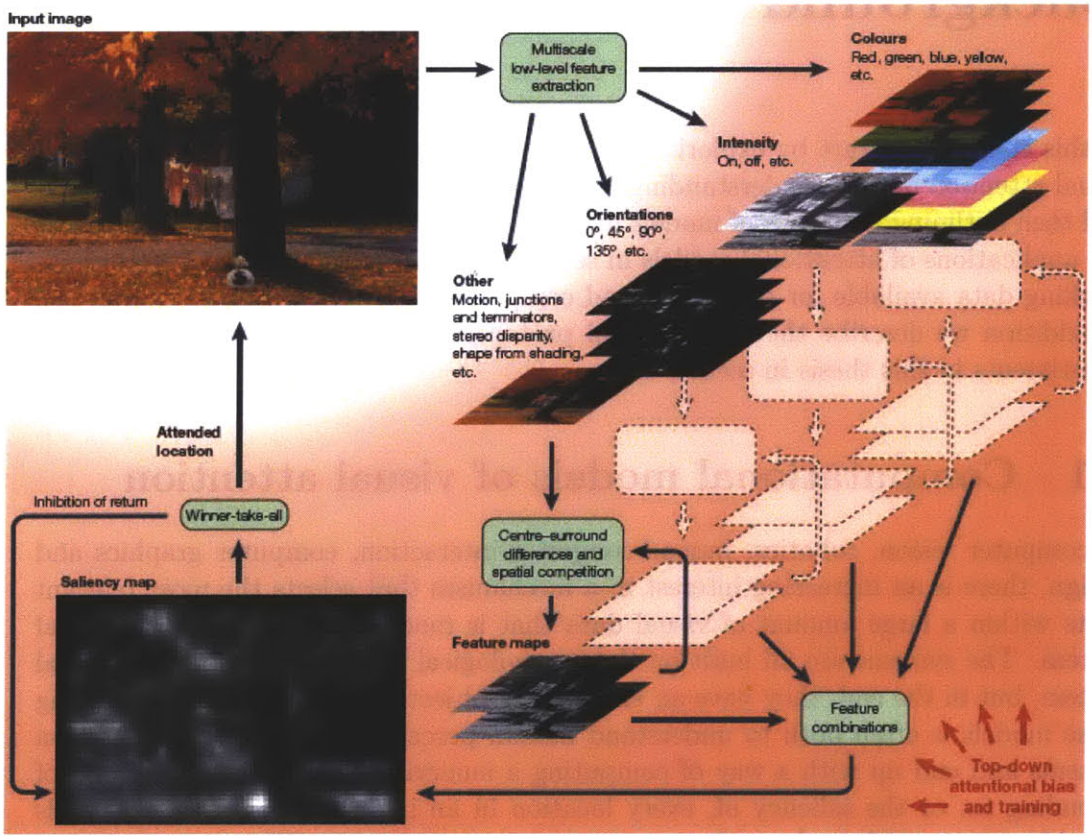

Figure 2-1: This is the basic structure of a feature-based computation model of visual attention. From Itti and Koch [2001].

model computes one or several image pyramids from the input image to enable the computation of features at different scales. Then, image features are computed. Commonly used features are intensity, color, and orientation. Each feature channel is subdivided into several feature types (for example, r, g, b maps for color). Center-surround mechanisms or differences of Gaussians are used to collect within-map contrast into *feature maps*. The operation compares the average value of a center region to the average value of a surrounding region. The feature maps are summed up to feature dependent maps called *conspicuity maps*. Finally, the conspicuity maps are normalized, weighted and combined together to form the saliency map. The saliency map is usually visualized as gray-scale image in which the brightness of a pixel is proportional to its saliency.

This saliency map might be regarded as the output to the model for it gives the saliency for each region of a scene. However, some applications might require the trajectory of image regions – mimicking human fixations and saccades. The selected image regions are local maxima in the saliency map. They might be determined by a *winner-take-all* approach and implemented with a notion of *inhibition of return* that ensures that all maxima are examined and prevents the focus of attention from staying at the most saliency region.

An important aspect of attentional systems is the way the different maps are fused. It is not clear how mapping and fusing happens in the brain, and computational systems use different approaches. Usually, a weighting function is applied to each map before summing up the maps. The weighting function determines the importance of features.

Before weighted maps are summed, they are usually normalized. This is done to weed out differences between a priori not comparable modalities with different extraction mechanisms. Additionally, it prevents channels that have more feature maps to be weighted higher than others.

After weighting and normalizing, the maps are summed to create the saliency map. Linear summation of feature channels into the final saliency map remains the norm.

The structure described so far is purely bottom up. Despite the well-known significance of top-down cues, most models consider only bottom-up computations because they are easier to model. Including other knowledge in a top-down matter is inspired by the Guided Search model and the theory of Biased Competition [Desimone and Duncan, 1995] and has been the subject of more recent models of saliency. This is typically done by modulating the weights of the conspicuity maps before they are combined based on some top-down information about the scene or the task. Other ways of adding top-down information include adding context information, or faces, text and object detectors.

We explore specific examples of computational models of visual attention in the next section. We make note in the footnotes if the model has been implemented in code that is available for download off the web. These are the models that are readily available for use in applications.

2.1.2 Important computational systems

The first computational model of visual attention was introduced by Koch and Ullman [1985]. When first published, the model was not yet implemented but provided the algorithmic reasoning for later implementations. An important contribution of their work is the winner-take-all (WTA) approach.

Clark and Ferrier [1988] were among the first to implement an attention system based on the Koch-Ullman model. It contains feature maps, which are weighted and summed up to a saliency map. Another early model was introduced by Milanese [1993]. This work introduced concepts like conspicuity maps and feature computations based on center-surround mechanisms that are still used in models today.

Another derivative of the Koch-Ullman model is the C++ *Neuromophic Vision Toolkit (NVT)* [1] which is implemented and kept up to date by Itti and colleagues [Itti *et al.*, 1998], [Itti and Koch, 2000], [Itti and Koch, 2001], [Navalpakkam and Itti, 2006]. This toolkit introduces image pyramids for the feature computations, which enables efficient processing.

Many others have tested this toolkit and suggested improvements: Parkhurst *et al.* [2002] modified the basic model to account for falloff in visual sensitivity. They noticed that the drop in visual sensitivity as a function of eccentricity on stimulus salience was an important determiner of attention and incorporated it in their model. Draper and Lionelle [2005] introduced SAFE (selective attention as a front end) which modified the original approach such that it is more stable with respect to geometric transformations like translations, rotations, and reflections. Walther and Koch [2006] extended this NVT model to attend to proto-object regions and created Saliency-ToolBox (STB)[2]. Harel *et al.* [2007] exploit the power, structure and parallel nature of graph algorithms to achieve efficient saliency computations of their Graph Based Visual Saliency[3] model, which is based on the use of a *dissimilarity* metric. Le Meur et al. (2006)[Le Meur *et al.*, 2006] [Meur *et al.*, 2007a] adapted the Koch-Ullman model to include the features of contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions. Others have updated the Koch-Ullman model by adding features such as symmetry [Privitera and Stark, 2000] or curvedness [Valenti *et al.*, 2009].

One of the oldest attention models that is widely known and still developed further is Tsotsos' selective tuning model of visual attention [Tsotsos 1990; 1993; Tsotsos et al. 1995]. It consists of a pyramidal architecture with an inhibitory beam. The model has been implemented for several features including luminance, orientation, color componency [Tsotsos et al. 1995], motion [Tsotsos 2005] and depth from stereo vision [Bruce and Tsotsos 2005a]. Originally the selective tuning model processed one feature dimension only, but later it was extended to perform feature binding [Rothenstein and Tsotsos 2006b; Tsotsos et al. 2008].

The above approaches are based on biologically motivated feature selection, followed by center-surround operations, which highlight local gradients. Recently, some

[1] http://ilab.usc.edu

[2] http://www.saliencytoolbox.net

[3] http://www.klab.caltech.edu/ harel/share/gbvs.php

have hypothesized that fundamental quantities such as "surprise" and "self-information" and "signal to noise ratio" are at the heart of saliency and attention. Itti and Baldi [2006] introduced a Bayesian model of surprise that aims to predict eye movements. Bruce and Tsotsos [2006] [2009] present a novel model for visual saliency computation built on a first-principles information-theoretic formulation dubbed Attention based on Information Maximization (AIM)[4]. They model bottom-up saliency as the maximum information sampled from an image. More specifically, saliency is computed as Shannon's self-information $-\log p(f)$, where f is a local visual feature vector. Navalpakkam and Itti [2005], [2006], [2007] define visual saliency in terms of signal to noise ratio (SNR). The model learns the parameters of a linear combination of low level features that cause the highest expected SNR for discriminating a target from distractors.

Models that add top-down components The majority of the models described so far are bottom-up. However, it is well known that task is a strong influencer on our attention [Yarbus, 1967], especially in the context of search. In fact Henderson *et al.* [2007] provide evidence that top-down information dominates real-world image search processes, such that the influence of low-level salience information on search guidance is minimal and others show that context is very important [Torralba *et al.*, 2006] [Oliva *et al.*, 2003]. In order to correctly mimic the attention of humans, we have to successfully merge both bottom-up and top-down influences.

Context of the scene is useful for speeding up search and recognition (we tend to look at the street rather than the sky when searching for our car) and can be added to models. For example, Torralba *et al.*'s contextual guidance model [2006] combines low-level salience and scene context when guiding search. Areas of high salience within a selected global region are given higher weights on an activation map than those that fall outside of the selected global region. The contextual guidance model outperformed a purely salience-driven model in predicting human fixation locations in a search task. This research has since been updated by Ehinger *et al.* [2009].

Similar to the contextual guidance model, Zhang *et al.* [2008] and Kanan *et al.*'s Saliency Using Natural statistics (SUN)[5] model [2009] combines top-down and bottom-up information to guide eye movements during real-world image search tasks. However, unlike the contextual guidance model, SUN implements target features as the top-down component. SUN once again outperformed a salience-driven model in predicting human fixation positions during real-world image search.

Both the contextual guidance model and the SUN model found that combining two sources of guidance significantly improved their abilities to predict human fixation locations, suggesting that humans similarly combine information types to guide search.

Goferman *et al.* [2010] present context-aware saliency[6] which aims at detecting the image regions that represent the scene and not just the most salient object. In

[4]Code and images available at http://www-sop.inria.fr/members/Neil.Bruce/AIM.zip

[5]http://cseweb.ucsd.edu/ 6zhang/code/imagesaliency.zip

[6]http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html

addition to including low-level features such as contrast and color, they also consider global effects which suppress frequently occurring objects, they add a notion that visual forms may possess several centers of gravity, and they include detectors of human faces.

A second way to add top-down component to a model is to modulate the weights of the feature maps depending on the task at hand as originally explored by Wolfe *et al.*, [1989]. For example, if searching for a vertical green bottle, the model would increase the weights of the green and vertical orientation feature maps to allow those features to be attributed more saliency. In the salience map thus formed, all scene locations whose features are similar to the target become more salient and are more likely to draw attention. Navalpakkam and Itti [2010], Elazary and Itti [2010] and Gao *et al.* [2008] use this approach.

Elazary and Itti [2010] propose a model called SalBayes which denotes the marriage between both saliency and Bayesian modeling. At its core, the model learns the probability of an object's visual appearance having a range of values within a particular feature map. In a search task, the model influences the various feature maps by computing the probability of a given target object for each detector within a feature map. As a result, locations in the maps with the highest probability are searched first.

Marchesotti *et al.* [2009] use context by proposing a model for saliency detection based on the principle that images sharing global visual appearances are likely to share similar salience. Assuming that a large annotated image database is available, they retrieve the most similar images to the target image, build a simple classifier and use it to generate saliency maps. Their main application is image thumbnailing.

Similarly, Gao *et al.* [2008], Gao and Vasconcelos [2004] and Gao and Vasconcelos [2005] propose a unified model for top-down and bottom-up saliency as a classification problem. They first applied this model to object detection [Gao and Vasconcelos, 2005] in which a set of features are selected such that a class of interest is best discriminated from other classes, and saliency is defined as the weighted sum of features that are salient for that class. In [Gao *et al.*, 2008][7], they defined bottom-up saliency using the idea that pixel locations are salient if they are distinguished from their surroundings. They used difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the KullbackLeibler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region.

A third way to add top-down guidance to models is to incorporate the use of object detectors. The work of Cerf et al. [2007][2008a][2008b][2009] confirmed that faces and text strongly attract attention and showed that they were difficult to ignore even when doing so imposes a cost. They refined the Itti and Koch [1998] model by adding a conspicuity map indicating the location of faces and text and demonstrate that this significantly improves the ability to predict eye fixations in natural images. They provide a working model[8] which combines the saliency map computation of

[7]http://www.svcl.ucsd.edu/projects/discsaltd/

[8]http://www.klab.caltech.edu/ moran/fifadb/code.html

Itti and Koch model with the locations of faces based on the Viola Jones [2001] face detection algorithm. Additionally, [Einhäuser *et al.*, 2008b] showed that objects predict fixations better than early saliency. They add a human defined object-map to Itti and Koch model and show that fixations are predicted better by objects than by early saliency.

Fourier based models Hou and Zhang [2007] proposed a spectral residual approach[9] based on the Fourier transform. The spectral residual approach does not rely on parameters and detects saliency rapidly. The difference between the log spectrum of an image and its smoothed version is the spectral residual of the image. Wang and Li [2008] build on Hou and Zhang's approach by combining spectral residual for bottom-up analysis with features capturing similarity and continuity based on Gestalt principles. Guo and Zhang [2010] later point out that the phase spectrum, not the amplitude spectrum, of an image's Fourier transform that is key to calculating the location of salient areas. They propose a novel multiresolution spatiotemporal saliency detection model called "phase spectrum of quaternion Fourier transform" (PQFT)[10] to calculate the spatiotemporal saliency map of an image by its quaternion representation.

Region-based models Another fundamental difference between saliency models is whether they are feature-based or incorporate some local grouping and are then called region-based models. Region-based models are well suited for object segmentation tasks [Achanta *et al.*, 2008] [Achanta *et al.*, 2009] [Liu *et al.*, 2007][Avraham and Lindenbaum, 2010]. Achanta *et al.* [2008] [2009] present a method for salient region detection that exploits features of color and luminance and outputs full resolution saliency maps with well-defined boundaries of salient objects. Liu *et al.* [2007] suggest that saliency can be learned from manually labeled examples. They formulate salient object detection as an image segmentation problem, where they separate the salient object from the image background. They use features of multiscale contrast, center-surround histogram and color spatial-distribution. Avraham and Lindenbaum [2006] and [2010] propose extended saliency (or ESaliency) that uses a validated stochastic model to estimate the probability that an image part is of interest. They use a region-based method by starting with a rough grouping of image regions, and then select regions that are unique with respect to the whole global scene rather than having local contrast.

Models that learn parameters Most of the methods based on Gabor or Difference of Gaussian filter responses require many parameters such as the number of filters, type of filters, choice of the nonlinearities, and a proper normalization scheme. These methods tend to emphasize textured areas as being salient regardless of their context. A good alternative is to use non-parametric approaches or learn the free parameters using techniques of machine learning.

[9] http://www.klab.caltech.edu/ xhou/

[10] Code available from the author

Kienzle *et al.* [2006] proposed to learn a visual saliency model directly from human eyetracking data using a support vector machine (SVM). Though the approach is promising, they use ground truth data of eye tracking fixations on a small database of grey-scale image of natural scenes that have no particular salient object.

Seo and Milanfar [2009b] [2009a] use local regression kernels as features which fundamentally differ from conventional filter responses. They use a nonparametric kernel density estimation for these features, that results in a saliency map constructed from a local self-resemblance measure, indicating likelihood of saliency.

Zhao and Koch [2011] use a least square technique to learn the weights associated with a set of feature maps from subjects freely fixating natural scenes drawn from four different eye-tracking data sets. They find that the weights can be quite different for different data sets, but face and orientation channels are usually more important than color and intensity channels.

Models that include a center bias. Researchers have shown several times that eye fixations tend to be biased towards the center of an image [Parkhurst and Niebur, 2003], [Tatler, 2007] [Tatler and Vincent, 2009] [Tseng *et al.*, 2009], [Le Meur *et al.*, 2006], [Bruce and Tsotsos, 2009]. Despite this, only [Parkhurst and Niebur, 2003] and [Zhao and Koch, 2011] have implemented a model that incorporates this bias.

If the reader is interested in further information, we the reader to a nice summary in [Tsotsos *et al.*, 2005], a review by Rothenstein and Tsotsos [2008] which presents a classification of models with details on the functional elements each includes, an overview by Shipp [2004] that compares different models along the dimension of how they map onto system level circuits in the brain and a nice survey by Frintrop *et al.* [2010]which covers models up to about 2005. However, the state of the art has changed dramatically in the last five years, warranting a new overview of the field.

2.2 Applications of models of visual attention

In situations where one needs to use large amounts of visual data, it is necessary to prioritize the information before processing. To do this, the concept of selective attention arouses much interest in data-intense fields such as computer vision, human-computer interaction, robotics, computer graphics. Selective attention provides an intuitive way to select interesting regions of an image in a natural way, and is a promising avenue to improve computer vision systems and graphics applications.

In addition, we can build smarter applications and more visually pleasing content by using models of visual attention that give us an idea of where people look. Images can be compressed more, websites can be designed better, street signs can be made safer, images can be thumbnailed better if we know what is likely to be important to a human and attract their attention.

We elaborate further on some specific applications for models of visual attention.

Computer Vision Detecting regions of interest (ROI) is an important method for some computer vision tasks. In image segmentation, parts of an image are grouped

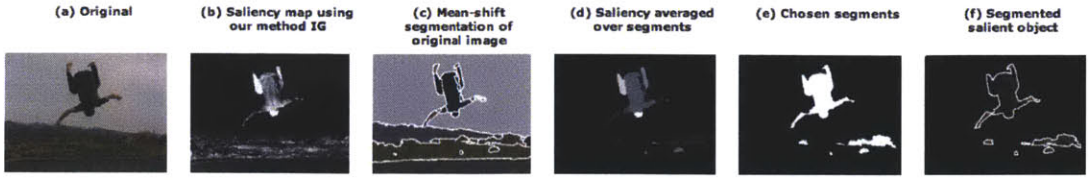

Figure 2-2: Saliency maps can help with image segmentation. Reprinted with permission from Achanta [Achanta *et al.*, 2009].

together depending on a measure of similarity by setting starting points (seeds) for the segmentation and choosing a similarity criterion. Salient regions are natural candidates for seeds and the homogeneity criterion can be adapted according to the features that discriminate a region from its surroundings. The model of saliency by Achanta *et al.* [2008] [2009] was designed specifically to help with the task ove object segmentation (see Fig. 2-2) as was the model of [Valenti *et al.*, 2009]. [Liu *et al.*, 2007] specifically formulates saliency as a process to detect salient objects.

Image compression algorithms can be enhanced by compressing regions that are not deemed important more than regions that are attended to. Given that saliency maps highlight regions considered important by humans, compression algorithms can adaptively determine the number of bits allocated for coding image regions according to their saliency. [Ouerhani, 2003] performs *focused image compression* with a visual attention system and [Itti, 2004] uses his saliency model to for video compression by blurring every frame increasingly with distance from salient locations. Guo and Zhang 2010[Guo and Zhang, 2010] A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression

Bottom up image saliency can be used to improve object recognition. [Miau *et al.*, 2001] present an object recognition approach that combines an attentional front-end with the biologically motivated object recognition system HMAX [Riesenhuber and Poggio, 1999]. [Walther and Koch, 2006] use a similar technique to combine an attentional system with a object recognizer based on SIFT [Lowe, 2004] and and show that recognition results improve with the attentional front-end. These systems rely on bottom-up saliency infomation and therefore assume that objects of interest are sufficiently salient by themselves. This works especially well for some object classes like traffic signs which are designed intentionally salient.

In the above approaches, the attentional part is separate from the recognition. However, in human perception these processes are strongly intertwined. A few groups have started on approaches in which both processes share resources. [Walther and Koch, 2006] suggest a unifying framework where the HMAX model for recognition is modulated to suppress or enhance locations and features due to spatial attention. A similar approach is used by [Navalpakkam and Itti, 2005] [Navalpakkam and Itti, 2007] [Elazary and Itti, 2010] and [Avraham and Lindenbaum, 2010]. Additionally, the attentional and object recognition processes are being brought together by saliency models that include object detectors. Some models [Cerf *et al.*, 2007], [Judd *et al.*,

2009], include Viola Jones [2001] face detectors or Felzenshwalb [2008] person, car and other object detectors.

Robotics In the process of robot localization, a robot has to determine its position by interpreting its sensor data and matching it to a given map of its surroundings. When the standard approaches of using laser scanners fail in outdoor environments, detecting and matching visual landmarks with a known position is a promising approach. Attentional models can facilitate the search of landmarks by selecting interesting regions in the sensor data. Siagian and Itti [2009] use salient objects in the environment as navigational landmarks. The system VOCUS of Frintrop allows for search of target objects [Frintrop, 2006], is largely robust to illumination and viewpoint changes and is capable of running in realtime [Frintrop *et al.*, 2007].

Models of attention can also be used to guide robot action. Robots, like humans, have to decide what to do next given their current state in the world. Part of this problem that is based on visual data is active vision, or deciding where to look next. One needs to direct the robotic camera to regions of potential interest in the same way that the human visual system directs gaze. In these applications, an attentional model which highlights the most salient features in the video stream are imperative. One of the first approaches to realize an active vision system with visual attention was done so by [Clark and Ferrier, 1988]. They describe how to direct a binocular robotic head with visual attention and perform simple experiments to fixate and track the most salient region in articial scenes. [Mertsching *et al.*, 1999] use the neural active vision system NAVIS on two different camera heads. Newer systems used for robot vision include [Zhang *et al.*, 2008] [Butko *et al.*, 2008].

Computer Graphics Within the field of computer graphics, a good saliency model can be used to provide suggestions of how to compute the best crop for a given size, as demonstrated by the initial work of [Santella *et al.*, 2006] in Fig. 2-3. They present an interactive method that uses fixation data provided by eye-tracking to identify important content and compute the best crop for any given aspect ratio or size. Suh *et al.* [2003] and Chen *et al.* [2003] built automatic image-cropping techniques that require no user input. Both systems identify important image areas using bottom up saliency models and face detection. These types of systems enable applications such as automatic snapshot recomposition, adaptive documents, and thumbnailing [Le Meur *et al.*, 2006] [Marchesotti *et al.*, 2009].

Saliency data used to determine regions of interest in a scene can help determine the level of detail appropriate to stylize and abstract photographs to make them more understandable, as seen in the work of [Grabli *et al.*, 2004] (see Fig. 2-4), or artistic [DeCarlo and Santella, 2002] (as in Fig. 2-5). A good measure of saliency can help suggest an appropriate threshold for the level of detail that a non photorealistic rendering of an image should include.

In content-aware media retargeting, an image or video is changed to fit a new aspect ratio such that it can be used on variable platforms and screen sizes. Selected pixels are deleted or morphed so that the resized image best portrays the important

(a) original (b) gaze-based (c) distracting crop

Figure 2-3: A good saliency method can help appropriately crop a photo. A good crop as produced by a gaze-based system for saliency can improve an image (b). A poorly chosen crop (c) can be distracting. Images from Santella *et al.* [2006].

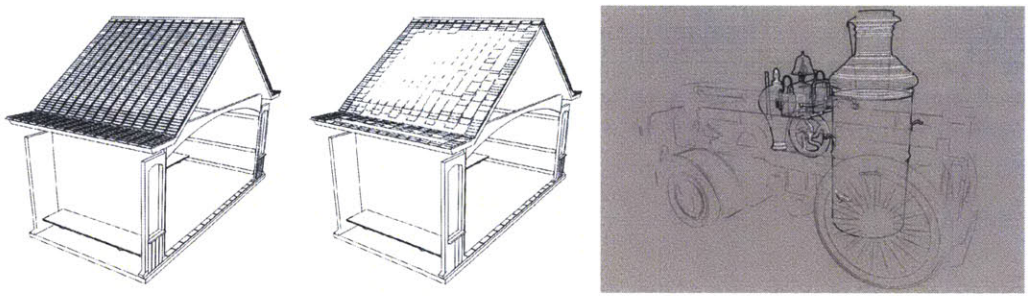

Figure 2-4: A saliency model can help determine the appropriate level and location of detail in the image. (Left) Two models of a house rendered at different levels of detail. (Right) An engine with a specific part rendered with more details. Images reprinted with permission from Grabli *et al.* [2004].

aspects of the original image, as in the work of [Avidan and Shamir, 2007], and [Rubinstein *et al.*, 2008]. All image retargeting methods use a cost function such as a saliency map to determine which pixels are the least important to prioritize them to be removed before important pixels (see Fig. 2-6) as done in the work of [Goferman *et al.*, 2010] and [Holtzman-Gazit *et al.*, 2010].

Design and marketing Companies are interested in knowing how observers look at and interact with their websites, if they are able to find what they need and navigate effectively. Advertisers would like to measure effectiveness of print designs to be able to optimize their messages (see Fig. 2-7)[11].

Companies such as SMIvision[12] and GazeHawk[13] offer eye tracking services which allow companies to to see how visitors view their webpage, advertisement, or image.

[11]Sunsilk images from http://www.rockyfu.com/blog/eyes-drive-attention/

[12]http://www.smivision.com/

[13]http://www.gazehawk.com

Figure 2-5: A model of attention can be used for stylization and abstraction of images as demonstrated in the work of [DeCarlo and Santella, 2002]. An eye tracker determines where people look in an image. The data is used to drive the level of detail at each location in the non photorealistic rendering of the image: more details where people looked and less elsewhere. Images reprinted with permission from DeCarlo and Santella [2002].

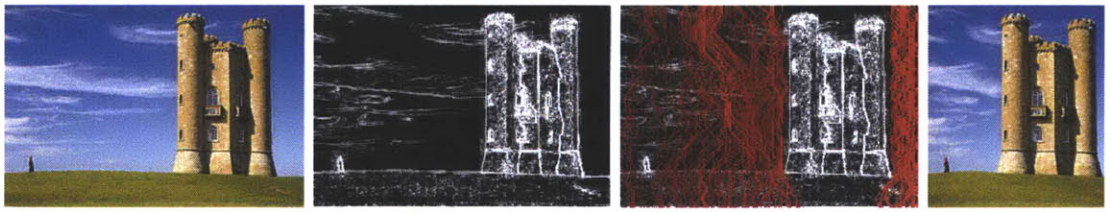

Figure 2-6: A saliency map of the original image on the left helps determine which seams of pixels in the image should be eliminated in order to produce the retargeting image on the right in the seam carving process of [Avidan and Shamir, 2007]. Images from wikipedia.

Figure 2-7: Eye tracking is used to aid design and market research.

Google does its own in-house eye tracking research[14]. If models of saliency become effective enough at modeling attention and predicting where people look, the need for expensive and intrusive eye tracking tests is reduced.

2.3 Data sets of fixations

In order to study the effectiveness of a saliency model to predict where people look in images, one needs a set of ground truth data of human fixations on images to compare against. To that end, researchers have run eye tracking experiments to create data sets of fixations on images. The data sets differ in the number and style of images chosen, the number of observers run, the exact task the observers were given (freeviewing, search, rating or memory task), but each help us understand where people actually look and measure performance of saliency models.

Recently the field has understood the importance of these data sets–many data sets have come out in the last two years. In this thesis, we introduce three new data sets. We use one to learn a new model of saliency and one to benchmark existing models of saliency. The third specialized data set is used to learn about fixations on low-resolution images.

Figure 2-8 shows eye tracking data sets available to the public as well as the datasets we introduce. Here is a description of each of the data sets in more detail.

Le Meur data set The data set published by Le Meur et al. [2006][15] has 27 color images with strongly salient objects. Every image was seen in random order by up to 40 observers for 15 seconds each in a task-free viewing mode. This set was originally used to assess the performance of the Le Meur [2006] et al model of saliency and compare it to Itti et al's [1998] model.

DOVES data set DOVES (a Database Of Visual Eye movementS)[16] is a collection of eye movements from 29 human observers as they viewed 101 natural calibrated images published by Linde et al.[2008]. The images contain only the

[14]http://googleblog.blogspot.com/2009/02/eye-tracking-studies-more-than-meets.html
[15]The data set is available at http://www.irisa.fr/temics/staff/lemeur/visualAttention/.
[16]The data set is available at http://live.ece.utexas.edu/research/doves/.

central 1024x768 pixels from the Natural Stimuli Collection created by Hans van Hateren. They are black and white and show natural outdoor scenes with no strongly salient objects. This database was published in order to encourage comparison of saliency models and the study of natural biases of human fixations.

FIFA data set For the data set[17] from Cerf et al.[2009], fixation data were collected from 8 subjects performing a 2-s-long free-viewing task on 180 color natural images (28 x 21). They were asked to rate, on a scale of 1 through 10, how interesting each image was. Scenes were indoor and outdoor still images in color. Images include salient objects and faces in various skin colors, age groups, gender, positions, and sizes. This data set was originally used to establish that human faces are very attractive to observers and to test models of saliency that included face detectors.

Itti and Baldi video data set Eye tracking data by [Itti and Baldi, 2006] was collected from eight subjects aged 2332 on a total of 50 video clips including indoor scenes, outdoor scenes, television clips, and video games. Video clips were displayed at a resolution of 640 x 480 and consist of over 25 minutes of playtime at approximately 60 Hz.

Toronto data set The data set from Bruce and Tsotsos[18] [2009] contains data from 11 subjects viewing 120 color images of outdoor and indoor scenes. Participants were given no particular instructions except to observe the images (32 x 24), 4 s each. One distinction between this data set and others is that a large portion of images here do not contain particular regions of interest. The data set was created to measure the performance of the authors' model of attention based on information maximization (AIM) against the Itti et al.[1998] model.

NUS data set The NUS data set[19] recently was introduced by Ramanathan et al. [2010] includes 758 images containing semantically affective objects/scenes such as expressive faces, nudes, unpleasant concepts, and interactive actions. Images are from Flickr, Photo.net, Google, and emotion-evoking IAPS [P.J. Lang and Cuthbert, 2008]. In total, 75 subjects free-viewed (26 x 19 or 1024x728 pixels) part of the image set for 5 s such that each image has an average of 25.3 subjects per image.

2.4 Experimental Protocol

In this thesis we perform and analyze the results of four separate eye tracking experiments. Though research questions differed, the experimental protocol for running the experiments stayed largely the same and is described below.

[17]The data set is available at http://www.fifadb.com/.

[18]The data set is available at http://www-sop.inria.fr/members/Neil.Bruce/.

[19]The data set is available at http://mmas.comp.nus.edu.sg/NUSEF.html.

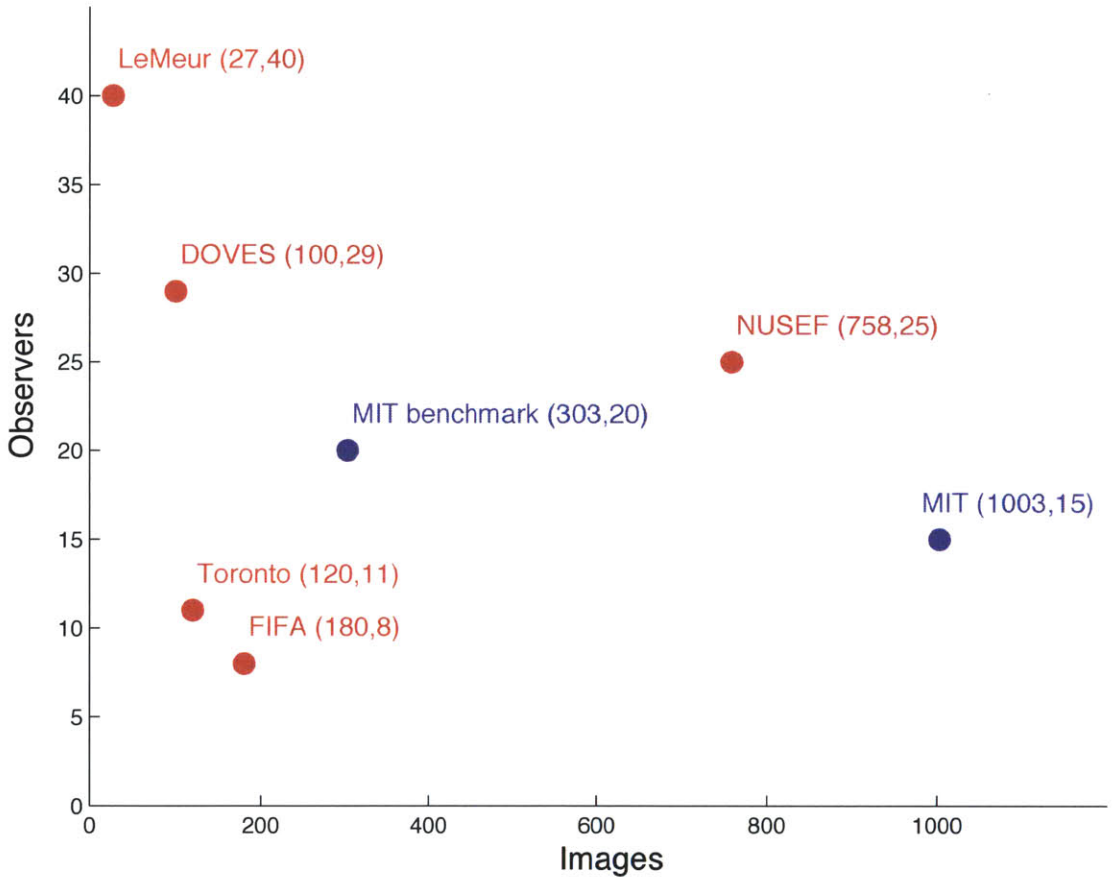

Figure 2-8: The chart shows different eye tracking data sets of natural full resolution images available (in red) compared to the databases we introduce (in blue). Our MIT training data set (1003, 15) is open and available to the public has a large number of images with eye tracking data that can be used to train new models of saliency. Our MIT benchmark dataset (300, 20) images are available for the public, but the observers' fixations are hidden and allow us to measure and compare the performance of different models.

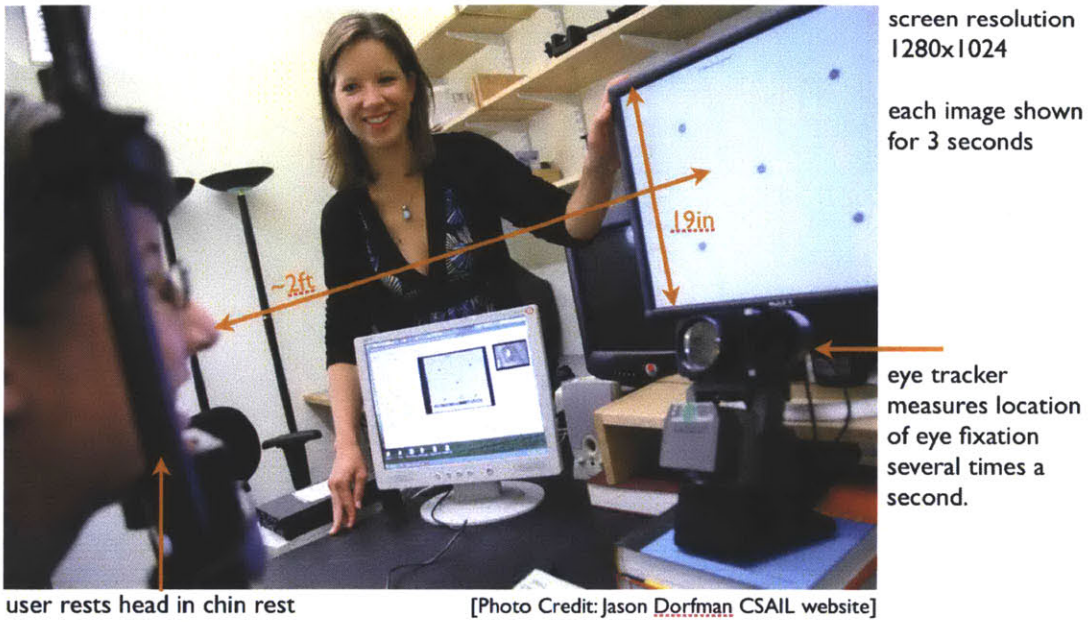

Figure 2-9: The setup for the eye tracking experiments.

For all of the eye tracking experiments, we used the same table-mounted, video-based ETL 400 ISCAN eye tracker which recorded observers' gaze paths at 240Hz as they viewed a series of images. All the viewers sat approximately 24 inches from a 19-inch computer screen of resolution 1280x1024px in a dark or semi-dark room and used a chin rest to stabilize their head. They typically saw a stream of between 36 to 1003 images broken into sections of reasonable length. When the observer and the eye tracker were well calibrated, the images auto-advanced and were usually shown for either 3 or 5 seconds.

We used a five point calibration system, during which the coordinates of the pupil and corneal reflection were recorded for positions in the center and each corner of the screen. We checked camera calibration either ever 50 or 100 images and recalibrated if necessary. The average calibration error was less than one degree of visual angle (35pixels). During the experiment, position data was transmitted from the eye tracking computer to the presentation computer so as to ensure that the observer fixated on a cross in the center of a gray screen for 500ms prior to the presentation of the next image.

The raw data from the eye tracker consisted of time and position values for each data sample. We used the method from [Torralba *et al.*, 2006] to define saccades by a combination of velocity and distance criteria. Eye movements smaller than the predetermined criteria were considered drift within a fixation. Individual fixation durations were computed as elapsed time between saccades and the position of each fixation was computed from the average position of each data point within the fixation. We discarded the first fixation from each scanpath to avoid the trivial information from the initial fixation in the center.

In some of the experiments, we motivated the observers to pay attention to the experiment by indicating that there would be a memory test at the end of the experiment. Independent of whether we ran the memory test or not, we did not use the data in our analysis.

Chapter 3

Learning to predict where people look

Abstract

Most saliency approaches are based on bottom-up computation that does not consider top-down image semantics and often does not match actual eye movements. To address this problem, we collected eye tracking data of 15 viewers on 1003 images and use this database as training and testing examples to learn a model of saliency based on low and high-level image features. The low-level features include the subbands of the steerable pyramid, colors, orientation, and intensity. High-level features include a horizon detector, and face, person and car detectors. We also include a center weighted feature that models the natural bias of human fixations to be near the center of the image. We show this combined model outperforms any model based on subsets of the features.

3.1 Introduction

Most models of saliency are biologically inspired and based on a bottom-up computational model that does not consider top-down image semantics. Though the models do well qualitatively, the models have limited use because they frequently do not match actual human saccades from eye-tracking data, as in Fig 3-2, and finding a closer match depends on tuning many design parameters.

We address these problems through two contributions in this chapter. The first is a large database of eye tracking experiments with labels and analysis, and the second is a supervised learning model of saliency which combines both bottom-up image-based saliency cues and top-down image semantic dependent cues. Our database consists of eye tracking data from 15 different users across 1003 images. To our knowledge, it is the first time such an extensive collection of eye tracking data is available for quantitative analysis. For a given image, the eye tracking data is used to create a "ground truth" saliency map which represents where viewers actually look (Fig 3-1). We propose a set of low, mid and high-level image features used to define salient

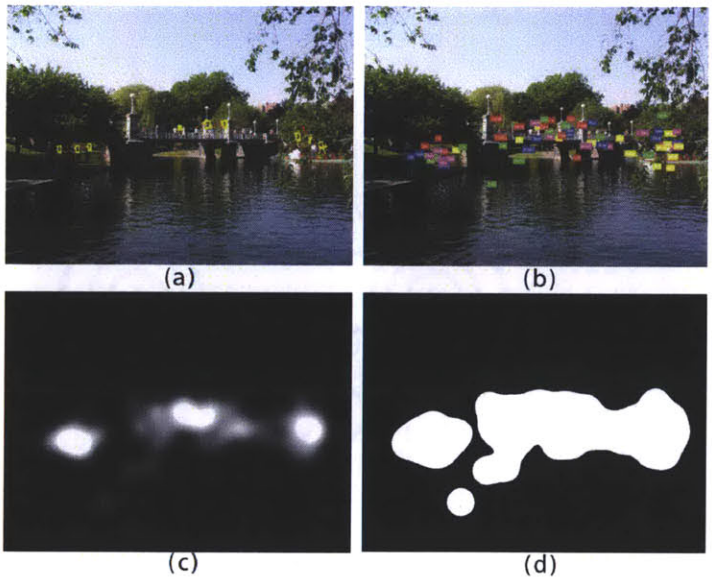

Figure 3-1: **Eye tracking data.** We collected eye-tracking data on 1003 images from 15 viewers to use as ground truth data to train a model of saliency using machine learning. Gaze tracking paths and fixation locations are recorded for each viewer (b). A continuous saliency map (c) is found by convolving a gaussian over the fixation locations of all users. This saliency map can be thresholded to show the most salient 20 percent of the image (d).

locations and use a linear support vector machine to train a model of saliency. We compare the performance of saliency models created with different features and show how combining all features produces the highest performing model. As a demonstration that our model can be used for graphics applications, we reimplement the DeCarlo and Santella [2002] abstracted nonphotorealistic rendering technique with our saliency model instead of eye tracking input.

Our work is most closely related to the work of Kienzle et al. [2006] who also learn a model of saliency directly from human eye movement data. Their model consists of a nonlinear mapping from a normalized image patch to a real value, trained to yield positive outputs on fixated patches, and negative outputs on randomly selected image patches. In contrast to our work, they only used low-level features. Furthermore, their training set comprises only 200 grayscale natural scene images.

3.2 Database of eye tracking data

We collected a large database of eye tracking data to allow large-scale quantitative analysis of fixations points and gaze paths and to provide ground truth data for saliency model research. The images, eye tracking data, and accompanying code in Matlab are all available on the web to facilitate research in perception and saliency across the vision and graphics community.

Figure 3-2: **Current saliency models do not accurately predict human xations.** *In row one, the low-level model selects bright spots of light as salient while viewers look at the human. In row two, the low level model selects the buildings strong edges and windows as salient while viewers fixate on the text.*

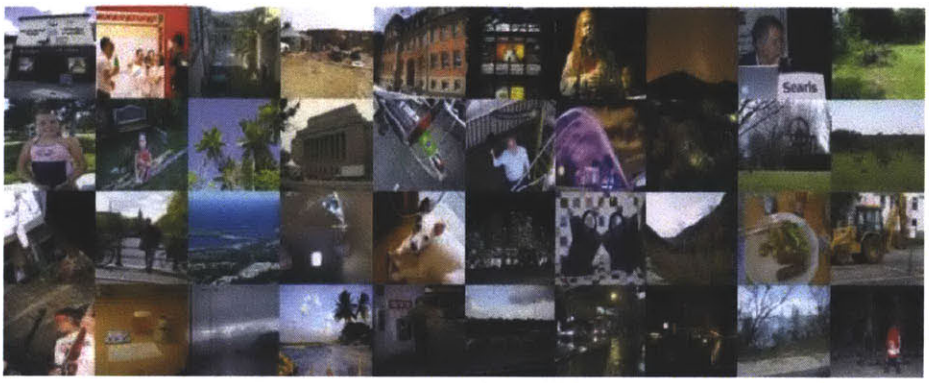

Figure 3-3: **Images.** *A sample of the 1003 images that we collected from Flickr and LabelMe. Though they were shown at original resolution and aspect ratio in the experiment, they have been resized for viewing here.*

3.2.1 Data gathering protocol

We collected 1003 random images from Flickr creative commons and LabelMe [Russell *et al.*, 2005] (Fig 3-3) and recorded eye tracking data from fifteen users who free viewed these images. The longest side of each image was 1024 pixels, and most images were 768x1024 or 1024x768 pixels in size though a few had a different aspect ratio. The users were both males and females between the ages of 18 and 35. Two of the viewers were researchers on the project and the others were naive viewers. All viewers sat at a distance of approximately two feet from a 19 inch computer screen of resolution 1280x1024 in a dark room and used a chin rest to stabilize their head. An eye tracker recorded their gaze path on a separate computer as they viewed each image at full resolution for 3 seconds separated by 1 second of viewing a gray screen. To ensure high-quality tracking results, we checked camera calibration every 50 images. We divided the viewing into two sessions of 500 randomly ordered images. Each session was done on average at one week apart. We provided a memory test at the end of both viewings to motivate users to pay attention to the images: we showed them 100 images and they had to indicate which ones they had seen before. We discarded the first fixation from each scanpath to avoid adding trivial information from the initial center fixation.

In order to obtain a continuous saliency map of an image from the eye tracking data of a user, we convolve a gaussian filter across the user's fixation locations, similar to the "landscape map" of [Velichkovsky *et al.*, 1996]. We also generate a saliency map of the average locations fixated by all viewers. We can choose to threshold this continuous saliency map to get a binary map of the top n percent salient locations of the image (Fig 3-1).

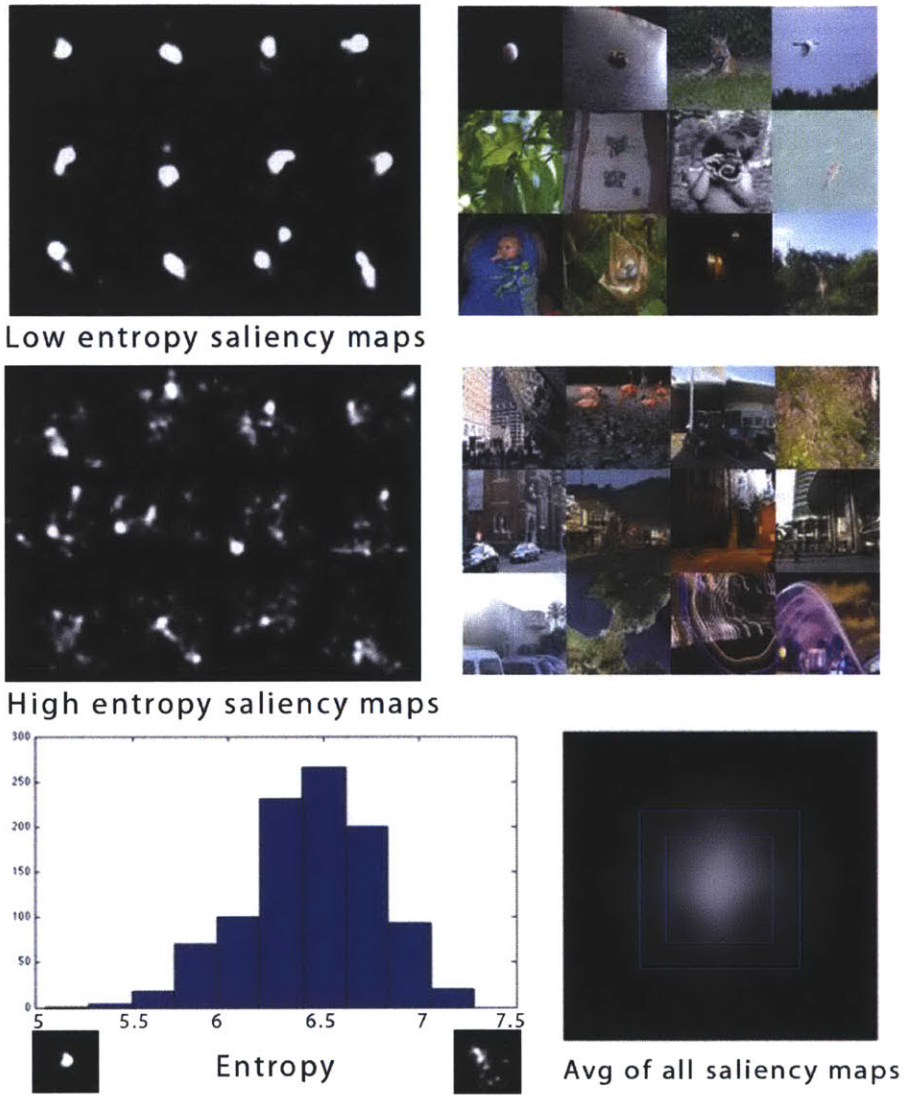

Figure 3-4: **Analysis of fixation locations.** *The first two rows show examples of saliency maps made from human fixations with low and high entropy and their corresponding images. Images with high consistency/low entropy tend to have one central object while images with low consistency/high entropy are often images with several different textures. Bottom left is a histogram of the saliency map entropies. Bottom right is a plot of all the saliency maps from human eye fixations indicating a strong bias to the center of the image. 40% and 70% of fixations lie within the indicated rectangles.*

3.2.2 Analysis of dataset

For some images, all viewers fixate on the same locations, while in other images viewers fixations are dispersed all over the image. We analyze this consistency of human fixations over an image by measuring the entropy of the average continuous saliency map across viewers. Though the original images were of varying aspect rations, we resized them to 200x200 pixel images before calculating entropy. Figure 3-4 shows a histogram of the entropies of the images in our database. It also shows a sample of 12 saliency maps with lowest and highest entropy and their corresponding images.

Our data indicates a strong bias for human fixations to be near the center of the image, as is consistent with previously analyzed eye tracking datasets [Zhang *et al.*, 2008]. Figure 3-4 shows the average human saliency map from all 1003 images. 40% of fixations lie within the center 11% of the image; 70% of fixations lie within the center 25% of the image. This bias has often been attributed to the setup of the experiment where users are placed centrally in front of the screen, and to the fact that human photographers tend to place objects of interest in the center of photographs [Zhang *et al.*, 2008].

We use an ROC metric to evaluate the performance of human saliency maps to predict eye fixations. Using this method, the saliency map from the fixation locations of one user is treated as a binary classifier on every pixel in the image. Saliency maps are thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated. The human fixations from the other 14 humans are used as ground truth. By varying the threshold, the ROC curve is drawn and the area under the curve indicates how well the saliency map from one user can predict the ground truth fixations. Figure 3-5 shows the average ROC curve over all users and all images. Note that human performance is remarkably good: 60% of the ground truth human fixations are within the top 5% salient areas of a novel viewer's saliency map, and 90 percent are within the top 20 percent salient locations.

As stated before, the fixations in the database have a strong bias towards the center. Because of this, we find that simply using a Gaussian blob centered in the middle of the image as the saliency map produces excellent results, as noted for other datasets as well by [Zhang *et al.*, 2008] [Meur *et al.*, 2007b]. We plot the ROC curve for the center Gaussian on figure 3-5.

In order to analyze fixations on specific objects and image features we hand labeled our image dataset. For each image, we labeled bounding boxes around any faces and text, and indicated a line for the horizon if present. Using these labeled bounding boxes we calculated that 10% of fixations are on faces (Fig 3-6). Though we did not label all people, we noticed that many fixations landed on people (including representations of people like drawings or sculptures) even if their faces were not visible. In addition, 11% of fixations are on text. This may be because signs are innately designed to be salient (for example a stop sign or a store sign are created specifically to draw attention). We use these ground truth labels to study fixation prediction performance on faces and as a ground truth for face and horizon detection. We also qualitatively found that fixations from our database are often on animals,

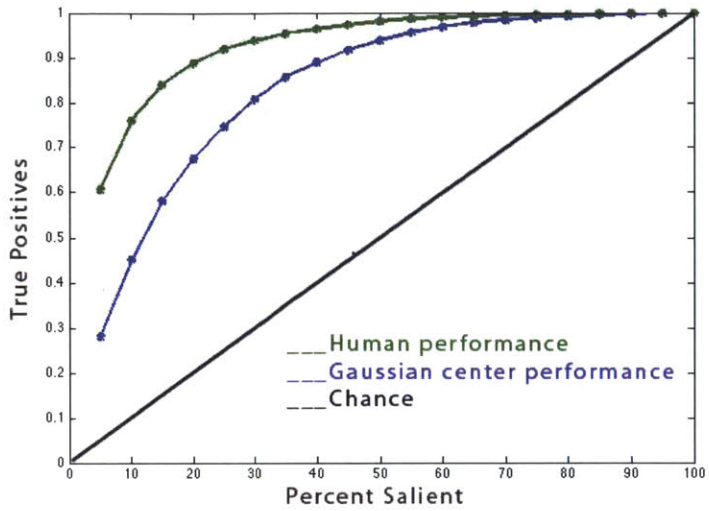

Figure 3-5: *In this ROC curve, human performance is very high demonstrating that the locations where a human looks are very indicative of where other humans have looked. The gaussian center model performs much better than chance because of the strong bias of the fixations in the database towards the center.*

cars, and human body parts like eyes and hands. These objects reflect both a notion of what humans are attracted to and what objects are in our dataset.

By analyzing images with faces we noticed that viewers fixate on faces when they are within a certain size of the image but fixate of parts of the face (eyes, nose, lips) when presented with a close up of a face (Fig 3-7). This suggests that there is a certain size for a region of interest (ROI) that a person fixates on. To get a quick sense of the size of ROIs, we drew a rough bounding box around clustered fixations on 30 images. Figure 3-7 shows the histogram of the radii of the resulting 102 ROIs. Investigating this concept is an interesting area of future work.

3.3 Learning a model of saliency

In contrast to previous computational models that combine a set of biologically plausible filters together to estimate visual saliency, we use a learning approach to train a classifier directly from human eye tracking data.

3.3.1 Features used for machine learning

The following are the low-, mid- and high-level features that we were motivated to work with after analyzing our dataset. For each image, we precomputed the features for every pixel of the image resized to 200x200 and used these to train our model.

Low-level features Because they are physiologically plausible and have been shown

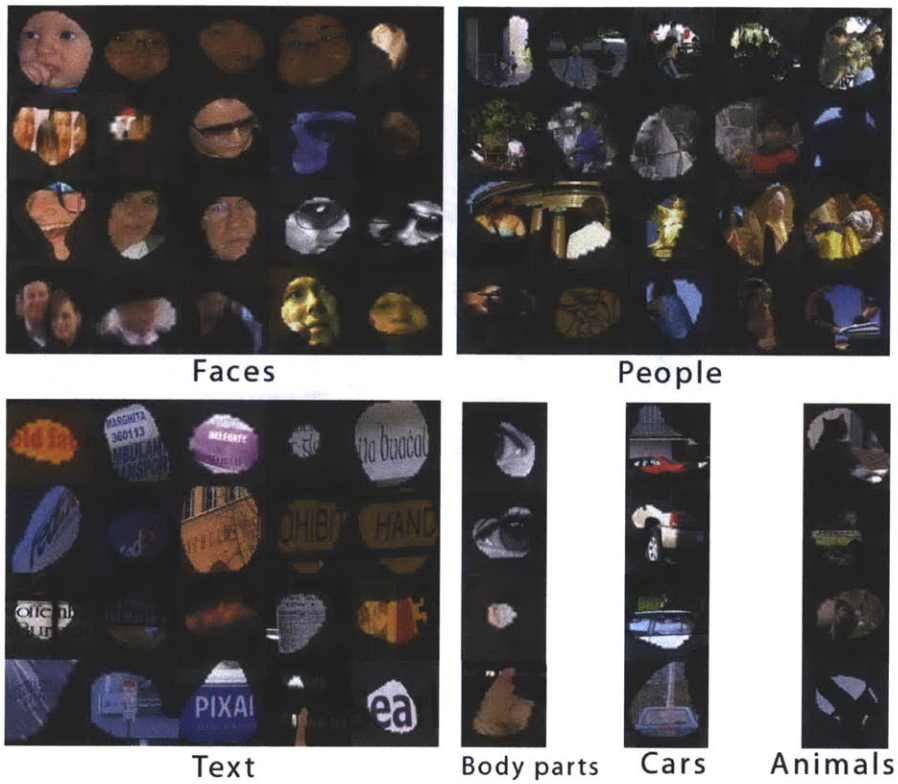

Figure 3-6: **Objects of interest.** *In our database, viewers frequently fixated on faces, people, and text. Other fixations were on body parts such as eyes and hands, cars and animals. We found these areas here by selecting bounding boxes around connected areas of salient pixels on an image overlayed with its 3% salient mask.*

Figure 3-7: **Size of regions of interest** *In many images, viewers fixate on human faces. However, when viewing the close up of a face, they look at specific parts of a face rather than the face as a whole, suggesting a constrained area of the region of interest. On the right is a histogram of the radii of the regions of interest as the percent of the whole image.*

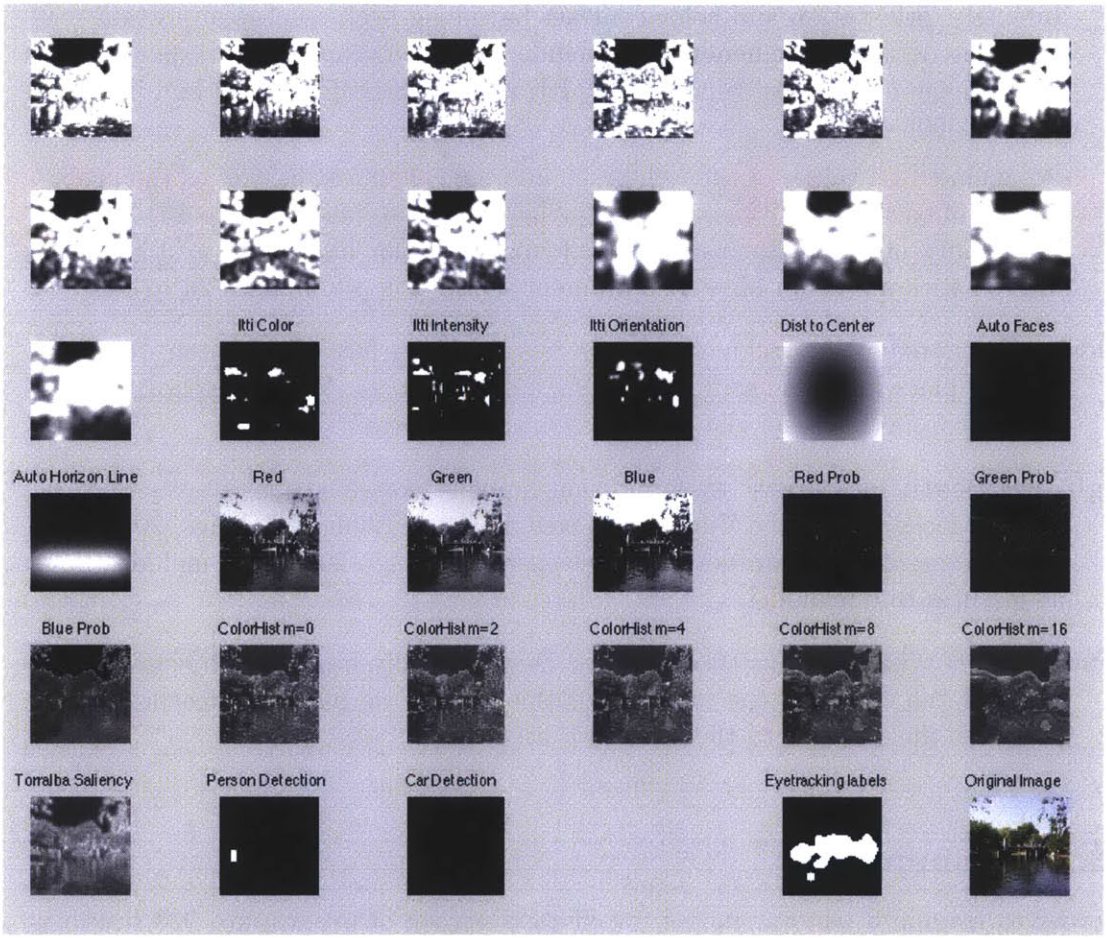

Figure 3-8: **Features.** *A sample image (bottom right) and 33 of the features that we use to train the model. These include subband features, Itti and Koch saliency channels, distance to the center, color features and automatic horizon, face, person and car detectors. The labels for our training on this image are based on a thresholded saliency map derived from human fixations (to the left of bottom right).*

to correlate with visual attention, we use the local energy of the steerable pyramid filters [Simoncelli and Freeman, 1995] as features. We currently find the pyramid subbands in four orientations and three scales (see Fig 3-8, first 13 images). We also include features used in a simple saliency model described by Torralba [Torralba *et al.*, 2006] and Rosenholtz [Rosenholtz, 1999] based on subband pyramids (Fig 3-8, bottom left).

Intensity, orientation and color contrast have long been seen as important features for bottom-up saliency. We include the three channels corresponding to these image features as calculated by Itti and Koch's saliency method [Itti and Koch, 2000].

We include the values of the red, green and blue channels, as well as the probabilities of each of these channels as features (Fig 3-8, images 20 to 25) and the probability of each color as computed from 3D color histograms of the image filtered with a median filter at 6 different scales (Fig 3-8, images 26 to 31).

Mid-level features Because most objects rest on the surface of the earth, the horizon is a place humans naturally look for salient objects. We train a horizon line detector from mid-level gist features [Oliva and Torralba, 2001a].

High-level features Because we found that humans fixated so consistently on people and faces we run the Viola Jones face detector [Viola and Jones, 2001] and the Felzenszwalb person detector [Felzenszwalb *et al.*, 2008] and include these as features to our model.

Center prior When humans take pictures, they naturally frame an object of interest near the center of the image. For this reason, we include a feature which indicates the distance to the center for each pixel.

3.3.2 Training

In order to train and test our model, we divided our set of images into 903 training images and 100 testing images. From each image we chose 10 positively labeled pixels randomly from the top 20% salient locations of the human ground truth saliency map and 10 negatively labeled pixels from the bottom 70% salient locations to yield a training set of 18060 examples and testing set of 2000 examples. We found that increasing the number of examples chosen per image above 10 did not increase performance. It is probable that after a certain number of examples per image, new examples only provide redundant information. We chose examples from the top 20% and bottom 70% in order to have examples that were strongly positive and strongly negative; we avoided examples on the boundary between the two. We did not choose any examples within 10 pixels of the boundary of the image.

We did find that the ratio of positive to negative training examples changed the performance of a given model by changing the threshold of how many positive and negative examples it returned: as the number of positive examples went up, the

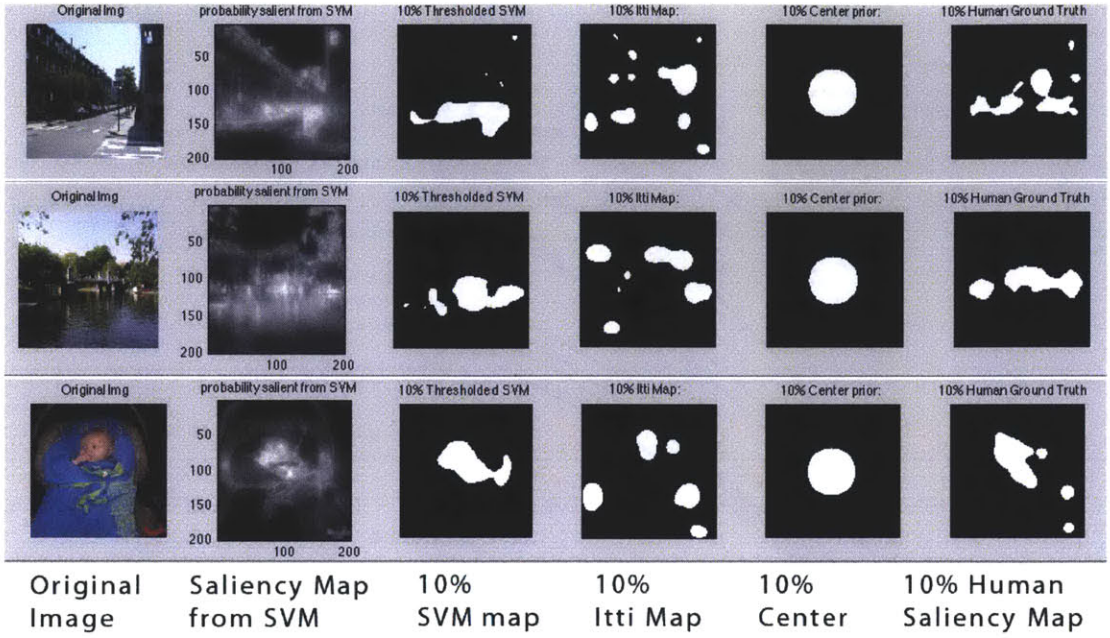

Figure 3-9: **Comparison of saliency maps.** *Each row of images compares the predictors of our SVM saliency model, the Itti saliency map, the center prior, and the human ground truth, all thresholded to show the top 10 percent salient locations.*

probability of returning a positive label went up as well. Because the ratio did not affect the *relative* performance between different models, we held the ratio constant at 1 to 1 to provide faster performance.

We normalized the features of our training set to have zero mean and unit variance and used the same normalization parameters to normalize our test data.

We used the liblinear support vector machine to train a model on the 9030 positive and 9030 negative training examples. We used models with linear kernels because we found from experimentation that they performed as well as models with radial basis function kernels and models found with multiple kernel learning [Sonnenburg *et al.*, 2006] for our specific task. Linear models are also faster to compute and the resulting weights of features are easier to understand. We set the misclassification cost c at 1. We found that performance was the same for $c = 1$, 10, 100 and decreased after that.

3.3.3 Performance

We measure performance of saliency models in two ways. First, we measure performance of each model by its ROC curve. Second, we examine the performance of different models on specific subsets of fixations: fixations inside and outside a central area of the image and on faces.

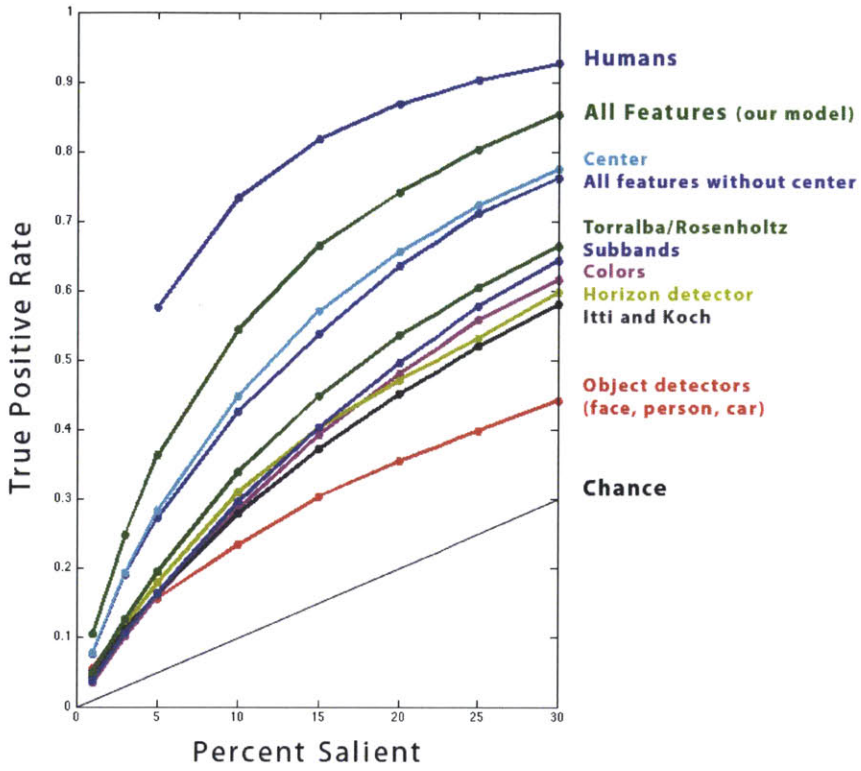

Figure 3-10: *The ROC curve of performances for SVMs trained on each set of features individually and combined together. We also plot human performance and chance for comparison.*

Performance on testing images In Figure 3-10, we see a ROC curve describing the performance of different saliency models averaged over all testing images. For each image we predict the saliency per pixel using a specific trained model. Instead of using the predicted labels (indicated by the sign of $w^T x + b$ where w and b are learned parameters and x refers to the feature vector), we use the value of $w^T x + b$ as a continuous saliency map which indicates how salient each pixel is. Then we threshold this saliency map at n =1, 3, 5, 10, 15, 20, 25, and 30 percent of the image for binary saliency maps which are typically relevant for applications. For each binary map, we find the percentage of human fixations within the salient areas of the map as the measure of performance. Notice that as the percentage of the image considered salient goes to 100%, the predictability, or percentage of human fixations within the salient locations also goes to 100%.

We make the following observations from the ROC curves: (1) The model with all features combined outperforms models trained on single sets of features and models trained on competing saliency features from Torralba and Rozenholtz and Itti and Koch. (2) The model with all features reaches 88% of the way to human performance.

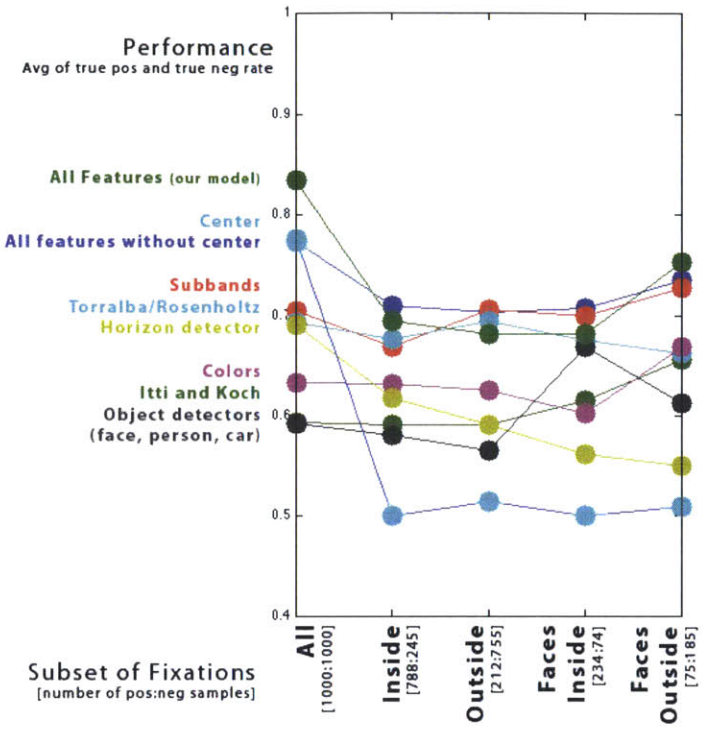

Figure 3-11: *Here we show the average rate of true positives and true negatives for SVMs trained with different feature sets on different* subsets *of fixations.*

For example, when images are thresholded at 20% salient, our model performs at 75% while humans are at 85%. (3) The model with all features except the distance to the center performs as well as the model based on the distance to the center. This is quite good considering this model does not leverage any of the information about location and thus does not at all benefit from the huge bias of fixations toward the center. (4) The model trained on all features except the center performs much better than any of the models trained on single sets of features. For example, at the 20% salient location threshold, the Torralba based model performs at 50% while the all-in-without-center model performs at 60% for a 20% jump in performance. (5) Though object detectors may be very good at locating salient objects when those objects are present in an image, it is not good at locating other salient locations when the objects are not present. Thus, the overall performance for the object detector model is low and these features should be used only in conjunction with other features. (6) All models perform significantly better than chance indicating that each of the features individually do have some power to predict salient locations.

Performance on testing fixations To understand the impact of the bias towards the center of the dataset for some models, we divided each image into a circular central and a peripheral region. The central region was defined by the model based only on

the feature which gave the distance of the example to the center. In this model, any fixation farther than 0.42 units away from the center (where the distance from the center to the corner is 1) was labeled negative and anything closer was labeled positive. This is equivalent to the center 27.7% of the image. Given this threshold, we divided the fixations to those *inside* and *outside* the center. In addition, we chose to look at fixations that landed on faces since viewers were particularly attracted by them.

In Figure 3-11 we plot performance of the model for different subsets of fixations. The performance here is defined as the average of the true positive and true negative rates.

We make the following observations about the trained models from this measure of performance: (1) Even though center model performs well over all the fixations (both fixations inside and outside the center), it performs only as well as chance for the other subsets of fixations. (2) While over all fixations the performance of the center model and the all-features-without-center model perform the same, later model performs more robustly over all subsets of images. (3) Understandably, the model trained on features from object detectors for faces, people and cars performs better on the fixations due to faces. (4) The SVMs using the center prior feature and the one using all features perform very well on 1000 positive and negative random testing points but are outperformed both in the inside and outside region. This paradox stems from the fact that 79% of the 1000 salient testing points are in the inside region, whereas 75% of the non-salient testing points are in the outside. One can show that this biased distribution provides a lift in performance for methods that would either have a high true negative rate outside or a high true positive rate inside, such as the center prior.

3.3.4 Applications

A good saliency model enables many applications that automatically take into account a notion of human perception: where humans look and what they are interested in. As an example, we use our model in conjunction with the technique of DeCarlo and Santella [2002] to automatically create a non photorealistic rendering of a photograph with different levels of detail (Fig 3-12). They render more details at the locations users fixated on and less detail in the rest of the image. While they require information from an eye tracking device in order to tailor the level of detail, we use our saliency model to predict locations where people look.

3.4 Conclusion

In this work we make the following contributions: We develop a collection of eye tracking data from 15 people across 1003 images which we will make public for research use. This is the largest eye tracking database of natural images that we are aware of and permits large-scale quantitative analysis of fixations points and gaze paths. We use machine learning to train a bottom-up, top-down model of saliency

Figure 3-12: **Stylization and abstraction of photographs** *DeCarlo and Santella [2002] use eye tracking data to decide how to render a photograph with differing levels of detail. More details are rendered at locations that humans fixate on. We replicate this application without the need for eye tracking hardware.*

based on low, mid and high-level image features. We demonstrate that our model outperforms the saliency model of Itti and Koch and the center prior. Finally, we show an example of how our model can be used in practice for graphics applications.

Discussion This eye tracking database allows us to quantify how consistent human fixations are across an image. In general, the fixation locations of several humans is strongly indicative of where a new viewer will look. So far, computer generated models have not matched humans' ability to predict fixation locations though we feel we have moved a step closer in that direction by using a model that combines both low and high level features.

Qualitatively, we learned that when free viewing images, humans consistently look at some common objects: They look at text, other people and specifically faces. If not people, they look at other living animals and specifically their faces. In the absence of specific objects or text, humans tend towards the center of the image or locations where low-level features are salient. As text, face, person and other object detectors get better, models of saliency which include object detectors will also get better. Though all these trend are not surprising, we are excited that this database will allow us to measure the trends quantitatively.

For future work we are interested in understanding the impact of framing, cropping and scaling images on fixations. We believe that the same image cropped at different sizes will lead viewers to fixate on different objects in the image and should be more carefully examined.

3.5 Lessons Learned

Overall we found that using more features gets better performance. Although considering more features usually results in more accurate and biologically plausible detection results, it also reduces the processing speed since the parallel models are usually

implemented sequentially. Therefore a trade-off has to be found between accuracy and speed.

One large open question related to this work is which features are most important? A full discussion of this is in the conclusion chapter of this thesis. The center feature was consistently the highest performing single feature. To get the ordering after that, we measured which features add most to our model by calculating the delta improvement between the center model and the center model with a given set of features. We observe that subband features and Torralbas feature map (which use subband features) add the greatest improvement. After that is color features, horizon detection, face and object detectors, and Itti channels.

The center bias of fixations in our data set is very strong. Borji *et al.* [2011] ask whether this is due to the photographic bias of our data set. They find that the data set does have a bias for objects near the center but also find that there is still a strong bias of fixations toward the center even on new, non-center biased data sets, suggesting that observers' view-strategy is still strongly center biased. This agrees with the work of Tseng *et al.* [2009] that says that both the photographic bias and the viewing strategy are equally important in creating fixations patterns towards the center. This means that good models of fixations should continue to include this bias as part of their model.

We notice that the face feature is very useful (it gets a high weight) even though the face detector has many false positives and negatives. This is partly due to the fact that observers have a strong tendency to look at faces. In fact 10% of fixations from our data set land on faces. In follow-up to our work, Zhao and Koch [2011] also learned weights for a four-feature model (color, orientation, intensity and faces) using our data set. They found that the weight for the face channel was highest, followed by orientation, color and intensity. The face weight was twice as high as the orientation weight. In our experiment, we also found that 11% of fixations landed on text. Adding a text detector would likely be an equally useful feature channel.

Related to the goal of finding which features work best is the goal of finding the *minimum* number of features that work. Our model includes 6 features based on color, 13 subbands of the steerable pyramid, and multiple other sets of features which possibly have much redundant information. It remains future work to cull down the number of features and still maintain high performance.

The work described in this chapter were published in [Judd *et al.*, 2009]. Since then, the technique of using machine learning to learn weights of features (instead of tuning them in an ad hoc fashion) has continued to gain momentum [Zhao and Koch, 2011], [Vig *et al.*, 2010].

Chapter 4

Benchmarking models of saliency

Abstract

There are many computational models of visual attention which are created from a wide variety of different approaches that aim to predict where people look in images. Each model is introduced by demonstrating performances on new images, and it is hard to make immediate objective comparisons between the models. To alleviate this problem, we propose a benchmark data set containing 300 natural images with eye tracking data from 20 viewers to compare the performance of models. We run 10 different models of saliency and calculate each model's performance at predicting ground truth fixations using three different metrics: a receiver operating characteristic, a similarity metric, and the Earth Mover's Distance. We describe performances here and online[1] and provide a way for people to submit new models for evalutation. Overall we find that the Judd and Graph-based visual saliency models perform best. In general, blurrier models, and models that include a notion of the center bias of fixations, perform well.

4.1 Introduction

Over the past few years, a number of high-quality computational models of attention have been developed and the state of the art is rapidly improving. Often when a new model is introduced, the authors make a good effort to compare it to at least one or two or three state-of-the-art models and see how each performs on a set of images viewed by humans. This means that with each new model comes new comparisons with different models on new images. There is no clear way to quantitatively *compare all the models against each other.* To alleviate this problem, we propose a benchmark data set, containing 300 natural images with eye tracking data from 20 viewers, to compare the performance of all available models. Because models were initially inspired with the aim to mimic the human visual system, we propose that the best way to measure model performance is to see how well each performs at *predicting where people look* in images in a free-viewing condition.

[1] http://people.csail.mit.edu/tjudd/SaliencyBenchmark

Where people look in an image is affected both by bottom-up and top-down mechanisms of visual attention. Many models only model bottom-up mechanisms because it is hard to model the top-down mechanisms involved in all possible states of the observer (memories, culture, age, gender, experiences) and possible tasks (searching, browsing, recognizing). However, it is still a reasonable goal to try and model attention for an *average* observer with a *free viewing* task. By free viewing we mean to imitate situations in which viewers are observing their world without a specific goal. This means that the best saliency models would pick up intrinsically salient locations independent of the specific viewer or task.

Not only is this benchmark of images aimed to help us score and compare how well existing saliency models predict where people look, it is also meant to provide a way to score new models. It is no longer enough that models predict fixations "significantly above chance". To go further we need models that predict where people look very well and approach the performance of humans. The difference between the performance of the best models today and the performance of humans shows that there is still room for improvement.

4.2 Previous comparisons

Several authors have shown that regions of interest found by computational model of visual attention correlate with eye fixations [Privitera and Stark, 2000] [Parkhurst *et al.*, 2002] [Elazary and Itti, 2008] [Henderson *et al.*, 2007] [Ouerhani, 2003] [Bruce and Tsotsos, 2006] [Itti, 2005] [Peters *et al.*, 2005] and the reported area under receiver operating characteristic (ROC) curve, which measures how well the two correlate, has increased over time as models get better. However, it is unfair to compare these numbers directly as they come from different experiments using different images under different conditions.

To facilitate comparisons, several authors have produced openly available databases of images with fixations from eye tracking experiments [Le Meur *et al.*, 2006] [Linde *et al.*, 2008] [Cerf *et al.*, 2009] [Bruce and Tsotsos, 2006] [Judd *et al.*, 2009] [Ramanathan *et al.*, 2010] and executable code of models that others can use. We provide a description of these datasets in section 2.3 and provide links to these through our website.

Given the accessibility of image databases and code, some authors have recently made substantial comparisons. Zhang *et al.* [2008] use the area under the ROC curve and KL-divergence to measure performance and compare against methods by [Itti *et al.*, 1998], [Bruce and Tsotsos, 2006] and [Gao *et al.*, 2008]. Seo and Milanfar [2009a] use the same measurement protocol and compare their model against the same three above and the SUN model [Zhang *et al.*, 2008]. Zhao and Koch [2011] learn ideal weights for the Itti and Koch model using four different datasets. They use three different metrics, ROC curve, Normalized Scanpath Saliency (NSS) and Earth Mover's Distance, to evaluate the performance of the models.

Our work builds off of these comparisons. We differ in that we compare many more models (10 different ones) and that we compare them on a new data set of images with a large amount of viewers. We measure performance of models using the

area under the receiver operator characteristic (ROC), earth mover's distance (EMD) and a similarity metric. This helps us make objective statements about how well the models predict fixations but also how they are similar and different from each other.

4.3 Experimental design

4.3.1 Benchmark data set

Images We collected 300 images combined from Flickr Creative Commons and personal image collections and recorded eye tracking data from twenty users who free-viewed these images. The longest dimension of each image was 1024 pixels and the second dimension ranged from 457 to 1024 pixels with the majority at 768 pixels. There were 223 landscape images and 77 portrait images.

Observers 20 observers (age range 18-50 years) participated in our eye tracking study. Each reported normal or corrected-to-normal vision. They all signed a consent form and were paid $15 for their time.

Method We used a method similar to the eye tracking experiment described in section 2.4. We displayed each image for 3 seconds followed by a gray screen. We tested calibration every 100 images. The instructions for this test were "You will see a series of 300 images. Look closely at each image. After viewing the images you will have a memory test: you will be asked to identify whether or not you have seen particular images before." We stated that there was a memory test to motivate users to pay attention however, we did not run a memory test at the end. Observers are typically more attentive if they are given a task, and the memory task is considered the lightest and least intrusive task available that still mimics free-viewing conditions.

We obtain a continuous *fixation map* for an image from the eye tracking data by convolving a gaussian filter across fixation locations of all observers. Figure 4-1 shows 5 images from our dataset, the fixations of 20 viewers on the images, and the corresponding fixation maps. The fixation map is normalized to range between zero and one.

4.3.2 Saliency models

We compare ten computational models of attention in our benchmark. These models are major models of attention that have been introduced in the last five years (except for the Itti and Koch model which was originally implemented in 1998) and they offer executable code. In essence, these are the models that are *available and useable* to a user who might like to use one to build further applications.

We list the models we used here in the order they were introduced. The **Itti and Koch model** is based on the theoretical bottom-up feature-based model of Koch and Ullman [1985], and was later implemented and improved by [Itti *et al.*, 1998], [Itti and Koch, 2000], [Itti and Koch, 2001], [Walther and Koch, 2006]. We used two

Figure 4-1: Five images from our benchmark data set (top), the fixation locations from our 20 viewers (middle), and the corresponding fixation maps (bottom).

implementations of this model: one implemented in Walther's Saliency Toolbox[2] and another which comes in the GBVS package[3]. The **Graph Based Visual Saliency (GBVS) model**[4] [Harel *et al.*, 2007] is a graph based implementation of the Itti and Koch model based on the use of a dissimilarity metric. The **Torralba model**[5] [Torralba *et al.*, 2006] incorporates the context of the scene. **Hao and Zhang model**[6] [Hou and Zhang, 2007] is based on the spectral residual of an image in the Fourier domain. [Zhang *et al.*, 2008] proposed the **SUN saliency model**[7] using natural statistics based on a Bayesian framework to estimate the probability of a target at every location. **Achanta**[8] [Achanta *et al.*, 2009] provides a simple model which aims to cleanly extract objects from their background. Because the target application of this work is different from the current benchmark task, it is at a clear disadvantage. This model would be better assessed under an image segmentation task though it is still interesting to explore how it performs on our task. The **Bruce and Tsotsos model**[9] [Bruce and Tsotsos, 2009] is a model for visual saliency computation built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM). The **Judd model**[10] [Judd *et al.*, 2009] is introduced in this thesis and incorporates bottom-up and top-down image features and learns the appropriate weights for fusing feature channels together. The **Context-Aware**

[2]http://www.saliencytoolbox.net/

[3]http://www.klab.caltech.edu/ harel/share/gbvs.php

[4]http://www.klab.caltech.edu/ harel/share/gbvs.php

[5]Code provided by Antonio Torralba.

[6]http://www.its.caltech.edu/ xhou/projects/spectralResidual/spectralresidual.html

[7]http://cseweb.ucsd.edu/ l6zhang/

[8]http://ivrgwww.epfl.ch/supplementary_material/RK_CVPR09/index.html

[9]http://www-sop.inria.fr/members/Neil.Bruce/

[10]http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html

Saliency[11] [Goferman *et al.*, 2010] aims at detecting the image regions that represent the scene and not just the most salient object. Their goal is to find image regions that "tell the story of the image". For further descriptions of these models, refer to the related work in section 2.1 of this thesis.

Figure 4-2 shows saliency maps produced by each of the models. The models are shown from top to bottom in the following order: Judd, GBVS, Itti and Koch 2 (GBVS implementation), Itti and Koch (SaliencyToolbox implementation), Bruce and Tsotsos' AIM, Context Aware, Torralba, Hao and Zhang, SUN, and Achanta.

The models vary greatly in the amount of salient pixels they return. For example, notice the difference in white or salient pixels between the Itti and Kock model and the Bruce and Tsotsos AIM model. In order to make comparisons across them, we match the histograms of the saliency maps to the histogram of the human fixation map for each image. In essence, this equalizes the amount of salient pixels allowed for each image and asks each model to place the mass of salient pixels on the areas it finds most salient. Saliency maps with matched histograms are shown in Figure 4-3. We use these histogram-matched saliency maps for all our performance calculations although this change does not affect the performance for the ROC metric.

4.3.3 Baselines

In addition to comparing saliency models to each other, we compare them to three important baselines shown in Figure 4-4. They have been histogram-matched to the fixation maps so that each shows the same amount of salient pixels.

Chance This model randomly selects pixels as salient. This can be considered a lower bound for the performance of saliency models. No model should have lower performance than chance.

Center This model predicts that the center of the image is the most salient. As the distance from the center of the image to the pixel increases, the value of its saliency decreases. The model is created by stretching a symmetric Gaussian to fit the aspect ratio of a given image. This means that if the image is much longer than it is high, the gaussian will have a longer horizontal axis. This stretched Gaussian performs slightly better than an isotropic Gaussian because it accounts for the tendency of objects of interest to be spread along the longer accesss. Ideally, any good model of saliency should outperform the center model.

Human performance Humans should be the best predictors of where other humans will look. However they do not predict each other perfectly because of the variability of many human factors and the complexity of a given image. Under the ROC metric, we measure human performance as how well the fixation map from 19 observers predicts the fixations of the 20th observer. This is averaged across all observers for the final performance. The other metrics require different approaches for estimating human performance. In all metrics, human

[11] http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html

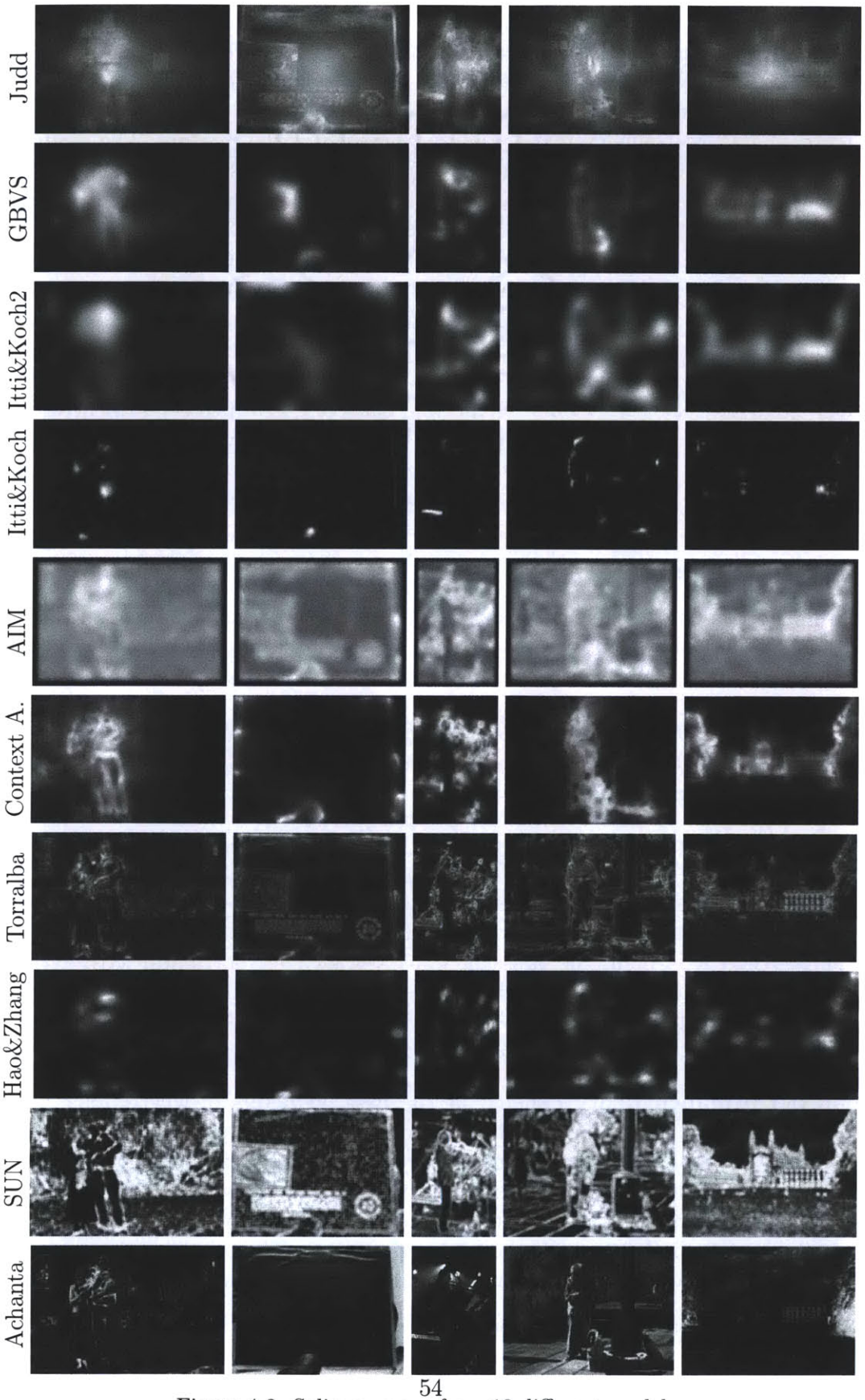

Figure 4-2: Saliency maps from 10 different models.

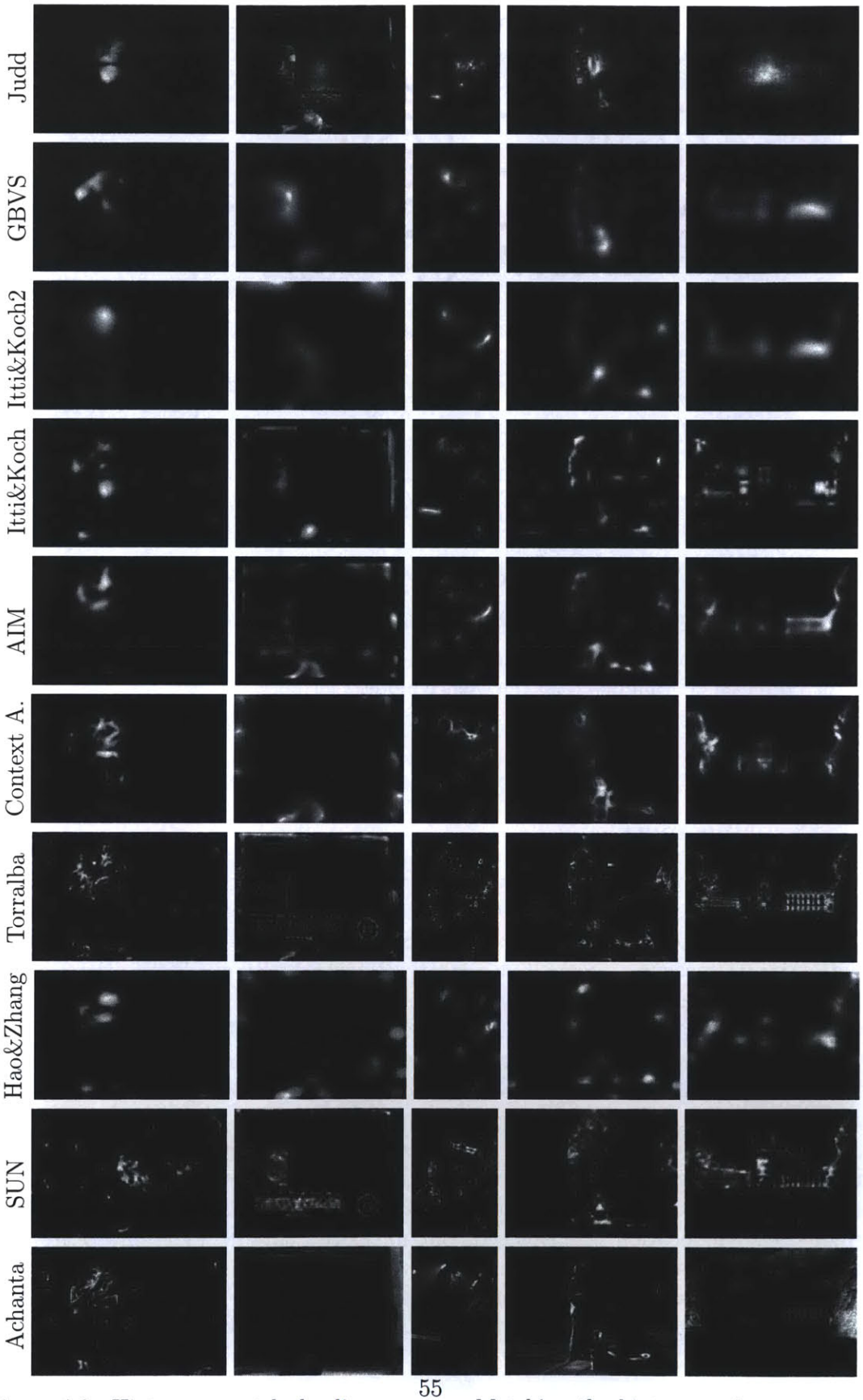

Figure 4-3: Histogram matched saliency maps. Matching the histograms between maps allows us to better understand which locations each model finds most salient

performance goes up as you increase the number of humans in a fixation map. Human performance provides the upper bound on the performance of saliency models: the best model possible would predict fixations of a new observer as well as other observers do.

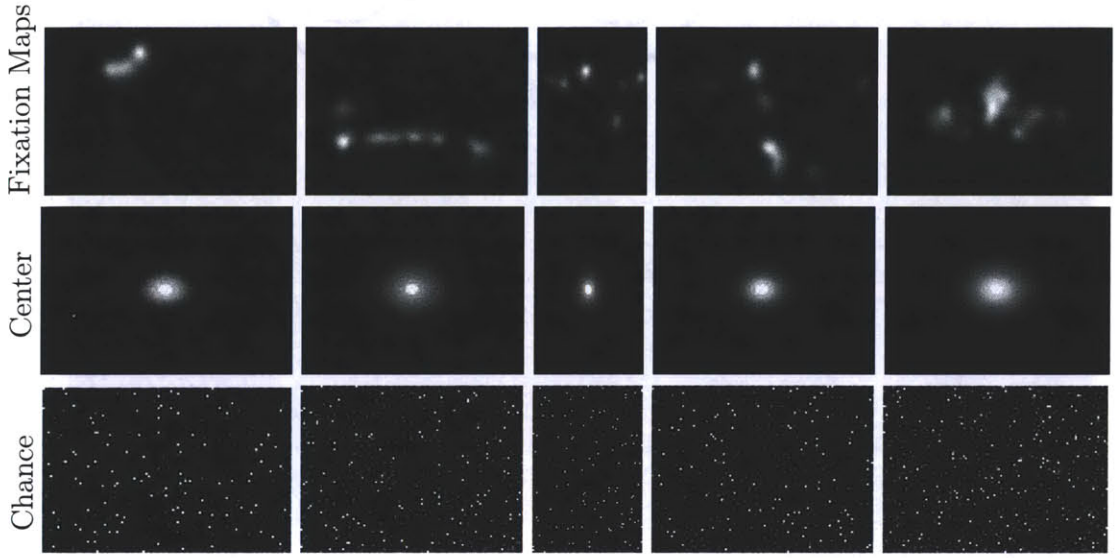

Figure 4-4: Baseline models for predicting where people look include human fixation maps (top), the center model (middle), and the chance model (bottom).

4.3.4 Scoring metrics

Several metrics can be used to quantitatively evaluate the performance of saliency models. These measures include the Receiver Operating Characteristics (ROC) [Green and Swets, 1966], the Normalized Scanpath Saliency (NSS) [Peters *et al.*, 2005], correlation-based measures [Jost *et al.*, 2005] [Rajashekar *et al.*, 2008], the least square index [Henderson *et al.*, 2007] [Mannan *et al.*, 1997], and the "string-edit distance [Brandt and Stark, 1997] [Choi *et al.*, 1995] [S. *et al.*, 1992].

Among these metrics, **ROC** is the most widely used in the community. According to Zhao and Koch [2011] he inherent limitation of ROC, however, is that it only depends on the ordering of the fixations. In practice, as long as the hit rates are high, the area under the ROC curve (AUC) is always high regardless of the false alarm rate. Therefore, while an ROC analysis is useful, it is insufficient to describe the spatial deviation of predicted saliency map from the actual fixation map. If a predicted salient location is misplaced, but misplaced close to or far away from the actual salient location, the performance should be different. To conduct a more comprehensive evaluation, we use a measure of similarity and the Earth Mover's Distance (EMD) [Rubner *et al.*, 2000] that measure the real spatial difference rather

than only the ordering of the values. Though the ROC performance is not affected by our histogram matching, the similarity and EMD performance is.

The **similarity score (S)** is a measure of how similar two distributions are. After each distribution is scaled to sum to one, the similarity is the sum of the minimum values at each point in the distributions. Mathematically, the similarity S between two maps A and B is

$$S = \sum_{i,j} \min(A_{i,j}, B_{i,j}) \text{ where } \sum_{i,j} A_{i,j} = \sum_{i,j} B_{i,j} = 1.$$

A similarity score of one means the distributions are the same. An similarity score of zero shows that they do not overlap at all and are completely different.

Earth Mover's Distance (EMD) [Rubner *et al.*, 2000] captures the global discrepancy of two distributions. Intuitively, given two distributions, EMD measures the least amount of work needed to move one distribution to map onto the other one. It is computed through linear programming and accommodates distribution alignments well. A larger EMD indicates a larger overall discrepancy between the two distributions. An EMD of zero indicates that two distributions are the same.

4.4 Experimental Results

In this section we show results of how well saliency maps predict human fixations based on the three metrics of performance ROC, similarity and EMD. We compare these to the performance of the baseline center and chance models and explain what human performance would be for each metric under different numbers of observers. We also examine which models have similar performance.

In addition to assessing models' performance at predicting where people look, we measure how similar the models are to each other. We create a similarity matrix to view relationships between models and plot the models in a higher dimensional space using multidimensional scaling.

Finally, we describe the complexity of our images based on how similar each observer's fixations are to the average of all observers fixations per image, and show how this image complexity affects saliency model performance results.

4.4.1 Model performances

Figure 4-5 shows the performance of saliency models using the three different metrics ROC, similarity and EMD. We measured performances using the histogram matched saliency maps from each model.

ROC The top chart of Figure 4-5 indicates how well saliency maps from each model predict ground truth fixations and shows the area under the ROC curve (AUR). For this metric, higher scores are better. We see that human performance is the highest. This provides an upper bound on possible performance of saliency maps. In addition, all models perform better than chance. The center baseline outperforms many models

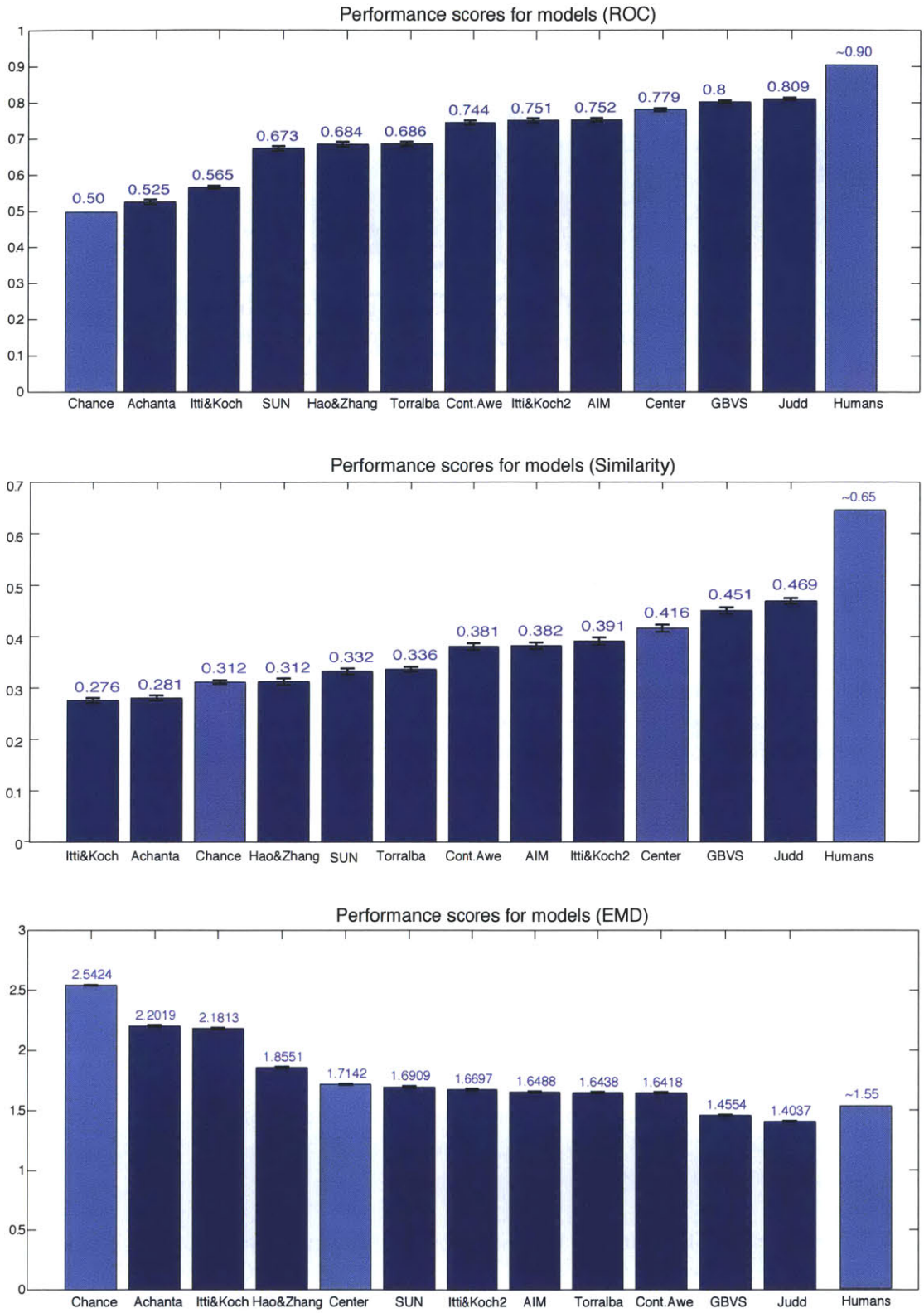

Figure 4-5: Performance of all models using the ROC, similarity, and EMD metric. For the first two graphs, higher values are better. Lower values are better for EMD performance. Error bars show the uncertainty of the mean over 300 images. The Judd and GBVS models perform the highest independent of the metric used.

of saliency. This is because of photographic bias of images and viewing-strategy of observers.

The Judd and GBVS are the highest performing models and the only models outperform the center on two of the three metrics. This is most likely because they are the only models to incorporate a global center bias to the model.

The Context Aware, AIM, and Itti and Koch2 models have about the same performance. Looking at the saliency maps of these models shows that they share similar visual properties and select similar salient locations. This is interesting since the models are fundamentally quite different: the Bruce and Tsotsos model aims to maximize information sampled from a scene and is derived directly from mathematical first principles. On the other hand, the Itti and Koch model is a feature-based model with several parameters and biologically inspired design choices. The context-aware model is based on the feature-based Itti and Koch model though it also includes global considerations, Gestalt principles and a face detector. Despite being made very differently, the models have similar performance. The models differ also in how blurry they are with Context Aware being the least blurry and Itti and Koch 2 being the most blurry. Another group of similar performing models includes Hao and Zhang, Torralba, and SUN models. Both Torralba and SUN saliency pick up small high frequency details in the image. Hou and Zhang locations are fundamentally different despite similar overall performance.

Itti and Koch and Achanta models perform the most poorly. Itti and Koch may perform poorly because its master map is not blurry. The Achanta model selects objects with clean boundaries, but they are not always, in fact often not, the salient objects. In general the Achanta model does very well on simpler images with one main salient object which the model selects very well from the background. When the images are complex scenes as in this benchmark, the model does not perform as well. For this data set, the model performs about as well as chance.

Finally, the two implementations of the Itti and Koch map perform very differently. This is mostly due to implementation details of the 2 methods used with their out-of-the-box parameters and demonstrates that selecting good parameters can have a very large effect on performance. In this particular case, Itti and Koch 2, which was implemented in the GBVS toolbox, has a larger level of blur applied to the master map.

Similarity The middle chart of Figure 4-5 represents how similar the saliency maps from different models are to the human fixation map. Higher scores indicate better performance and a perfect score of one would indicate that two maps are exactly the same. Overall the ordering of performances based on similarity is similar to the ordering from the ROC performances.

We reason about human performance in the following way: Similarity between two fixation maps that each have infinite viewers will approach a value of one because they'll start be be exactly the same. With only a finite number of viewers, fixation maps will always be slightly different as observers do not look at exactly the same places. We measure the similarity of human fixations maps when the number of

viewers per map is 2, 5, 8, and 10 to be 0.37, 0.52, 0.59 and 0.63 respectively. We extrapolate this on a linear log scale and predict that the similarity of two fixation maps each with 20 viewers would be around 0.65.

Earth Mover's Distance The bottom chart of Figure 4-5 measures how much mass needs to be moved to change the saliency map of a given model into the corresponding human fixation map. Lower values indicate better performance as less "earth" needs to be moved. Under this metric we see similar orderings for the best and worst models while the five mid-performing models and center all have about equal performance.

We obtain an estimate for the human performance in a similar way to that of the similarity metric. We estimate the EMD between two fixation maps created from 2, 5, 8, and 10 observers as 1.80, 1.84, 1.71, 1.67 respectively and extrapolate that to approximately 1.55 for 20 observers. As more observers are added to ground truth fixation maps, the EMD between two of them will decrease.

4.4.2 Multidimensional space analysis

To understand which models are similar to each other, we calculate the similarity between the saliency maps of each model on each image and average the similarities across all images. This gives us a similarity between each of the models which we plot as a similarity matrix in Figure 4-6 where the models are ranked according to their similarity to the human fixation maps. The squares with yellower or greener color represent models whose maps are more similar to each other; the bluer the square, the more different the maps are. The diagonals of the similarity matrix are red with value 1 because a saliency map of a given model is always the same as itself. The next highest value is the intersection of the Judd model and the Center baseline. This is because the Judd model includes a strong center bias. SUN saliency and the Torralba model are similar to each other once again. In addition, Bruce and Tsotsos AIM is similar to Hao and Zhang, Torralba, Context Aware and Itti and Koch2. Note that the center model is ranked quite high (fourth in the similarity matrix) yet most of its line is a very low scoring blue. This means that the other models are very dissimilar to it; most models do not include a notion of the center bias of fixations.

If we take 1 minus the similarity values, we get the dissimilarity values, or distances between models, that we can use to plot the models using multidimensional scaling. Using 2 dimensions accounts for 45% of the variance of the models, as per Figure 4-7, and is the 2D plot is seen in the top of Figure 4-8. Using 3 dimensions accounts for 60% of the variance and can be seen in the bottom of Figure 4-8. The closer the points representing the models are in 2D or 3D space, the more similar the models are. These plots make it apparent that the Itti and Koch model and the Achanta model are very different models from the rest; they are the outliers in the graph.

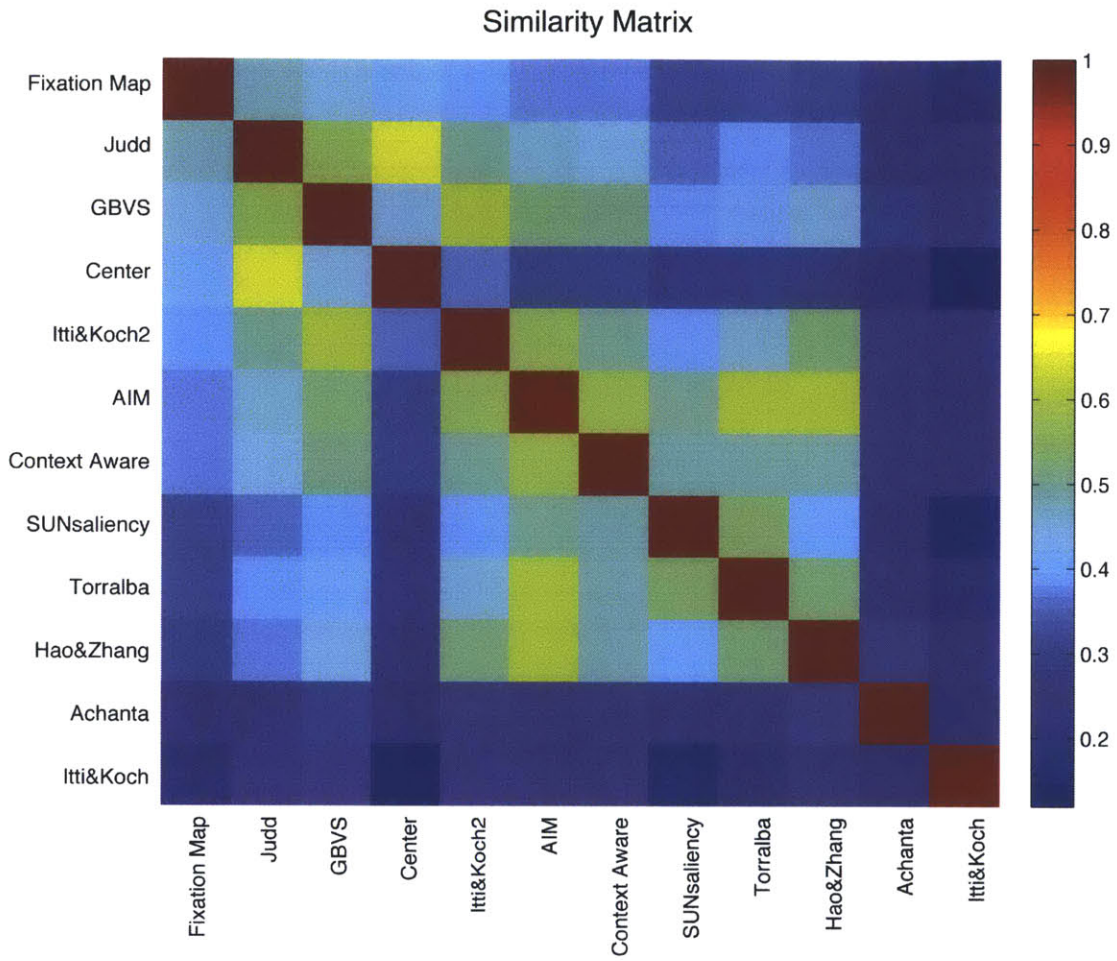

Figure 4-6: Similarity Matrix shows how similar each model is to all other models. They are plotted here by how similar they are to the fixation map.

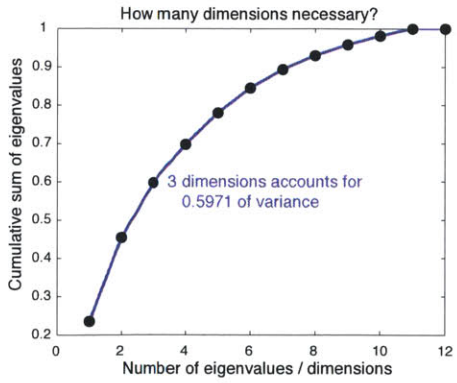

Figure 4-7: Variance of the models in mutlidimensional space

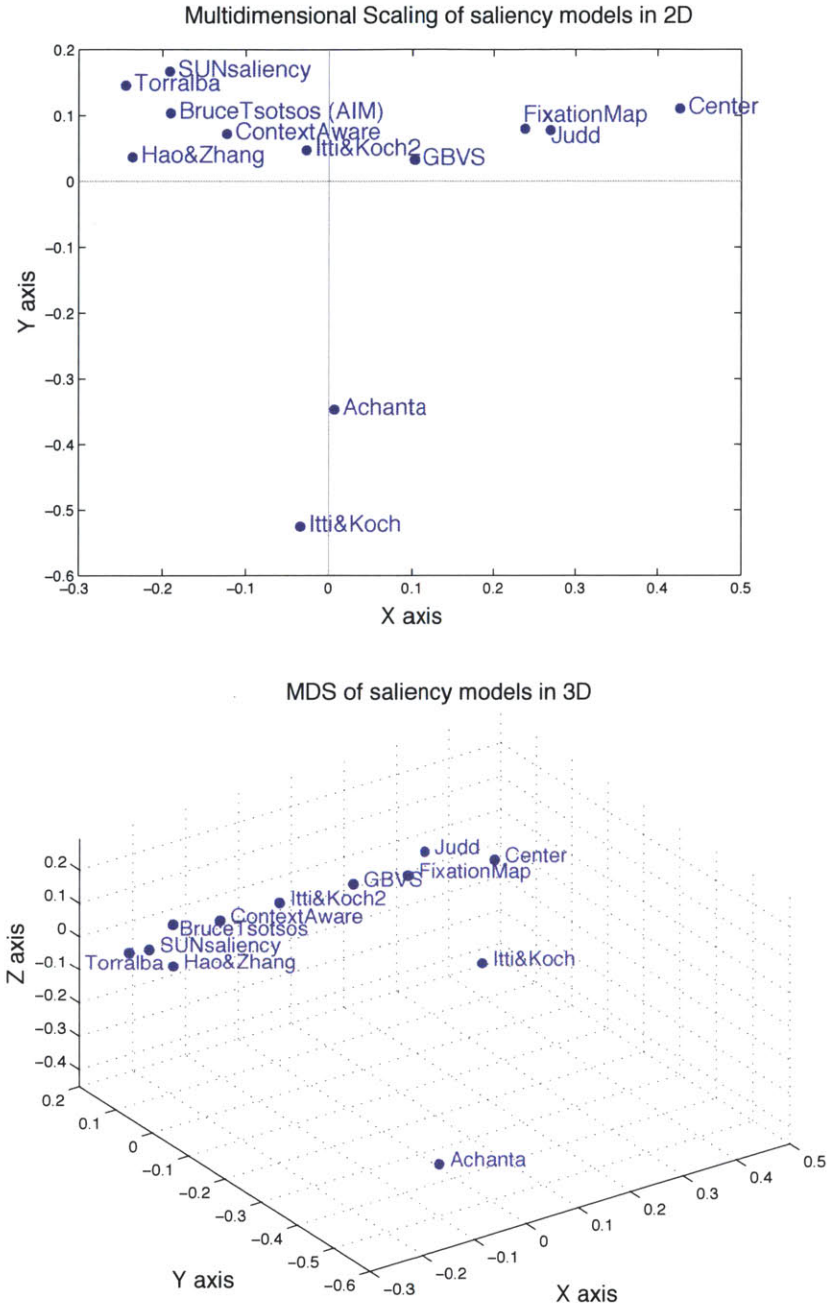

Figure 4-8: Models plotted in 2D space (top) and 3D space (bottom). The distance between the models's points gives an approximation of how separate the models are from each other in higher dimensional space.

4.4.3 Images ranked by fixation consistency

We note that on some images, all people look in the same few locations. In this case we say the human fixations are very *consistent*. We find the fixation consistency for each image by measuring the similarity between one observer's fixation map to the average fixation map of all observers. Images where all observers look at the same locations have very high similarity scores; images where observers fixations are very different from each other have low similarity scores. We rank images according to their consistency and the resulting order is seen in Figure 4-9 and in more detail in Figure 4-10. Images with people's faces, strong text, or one major object have have consistent fixations and high similarity scores. Complex outdoor scenes, cluttered data, or the notoriously detailed Where's Waldo images do not have consistent fixations and have low similarity scores. In these cases it may be harder for saliency models to accurately predict where people look. The ordering of the images according to this measure of fixation consistency is likely to be very similar to the order of the images according to their ROC performance.

Saliency model scores depend on image fixation consistency

To assess whether saliency model performance is affected by image fixation consistency, we used the image ranking from Figure 4-9 and divided the ranked images into three consecutive bins corresponding to high, medium and low fixation consistency. Figure 4-11 shows the performance of all models under all three metrics for images in each bin.

Note that as fixations consistency goes down, ROC scores decrease slightly, similarity scores decrease slightly, and earth mover's distances increase. This means that when humans are consistent it is easier for saliency models to predict where people will look; when human fixations are not as consistent with each other, saliency models have a harder time predicting fixations.

For future work, it would be interesting to see what the performance is for images with people vs without people, or landscapes vs objects, to understand if models are better at predicting fixations on certain types of images.

4.5 Online benchmark

In addition to the analysis provided here, we provide performance data of existing models and instructions for submitting new models for evaluation online at http://people.csail.mit.edu/tjudd/SaliencyBenchmark/.

The site includes a short description of each model, a link to the code for the model, and the performance with each metric. To allow others to understand the type of images and data in the data set, we include fixations and fixation maps for 10 of the images. We also include saliency maps from each model for direct visual comparison.

The website also includes instructions for submitting new models for evaluation. The instructions are summarized here: 1) Download all 300 images in this data set.

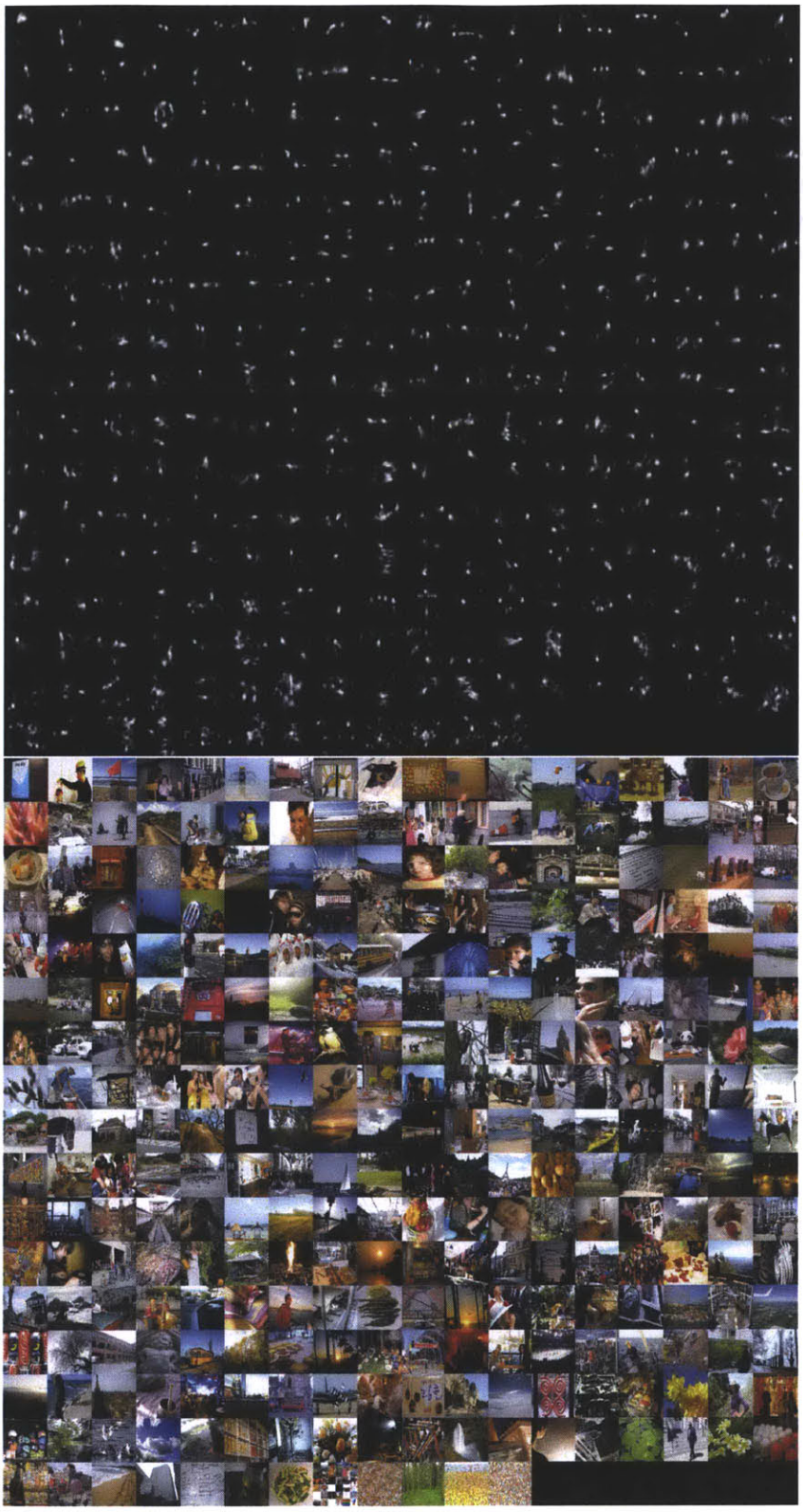

Figure 4-9: All images ranked by similarity of human fixations. Images where people look at the same places are shown first; images where people looked in several different places, or where fixations are spread across the image, are shown last.

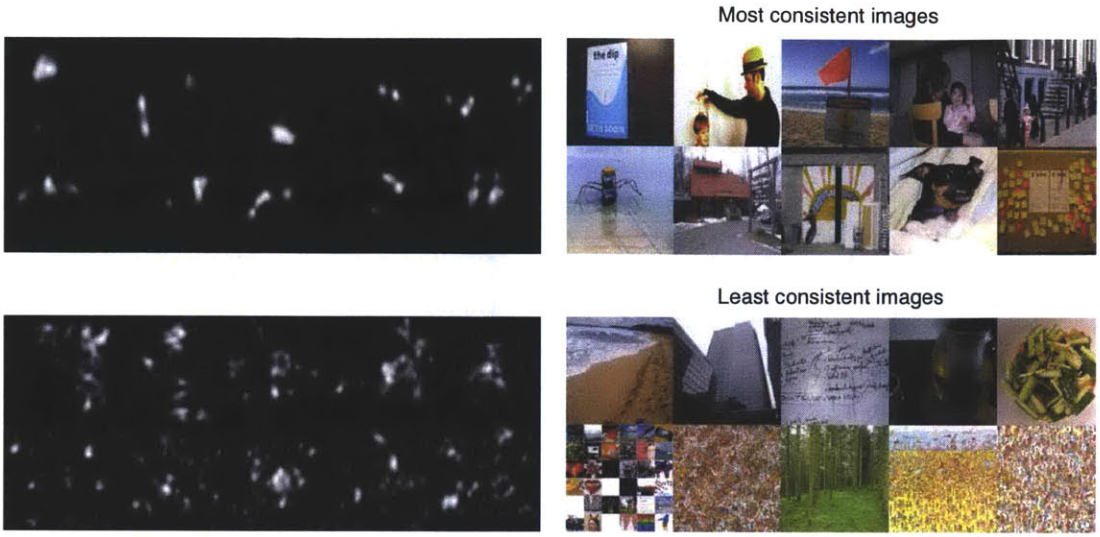

Figure 4-10: All images used ranked by similarity of human fixations. On the top are the images where humans are most consistent. The bottom row shows the fixation maps corresponding images where human fixations were most spread out.

2) Calculate saliency maps with the new model. The model should create maps that rank image locations by their likelihood to attract fixations of human viewers as it will be measured for its ability to predict where people look. 3) Submit the saliency maps via email to tjudd@csail.mit.edu or online using the submission interface. 4) Wait for your AUR, Similarity and EMD scores. As the actual fixations on the 300 images are hidden from the world to avoid training and over fitting models, we run the scoring calculations and send you the results. 5) Your scores and a description of your model will be included on the website.

If the reader is interested in training new models based on images with fixation data sets, we provide a link to our MIT dataset of 1003 images with 15 viewers introduced in chapter 3 and in [Judd *et al.*, 2009]. In addition we link to several other smaller data sets that are publicly available.

4.6 Conclusion

There are many models of saliency, including several that have been introduced in the last five years, but no large comparison of many of the current state of the art models on a standard data set. Our goal is to fill that need. We compare 10 recent models of saliency whose implementations are available online and measure how well they perform at predicting where people look in a free-viewing task under three different metrics. We found that the Judd [Judd *et al.*, 2009] and the Graph Based Visual Saliency models [Harel *et al.*, 2007] consistently outperform other models of saliency and the center model baseline. Other models have some predictive power as they perform better than chance but they often do not outperform the center model. The

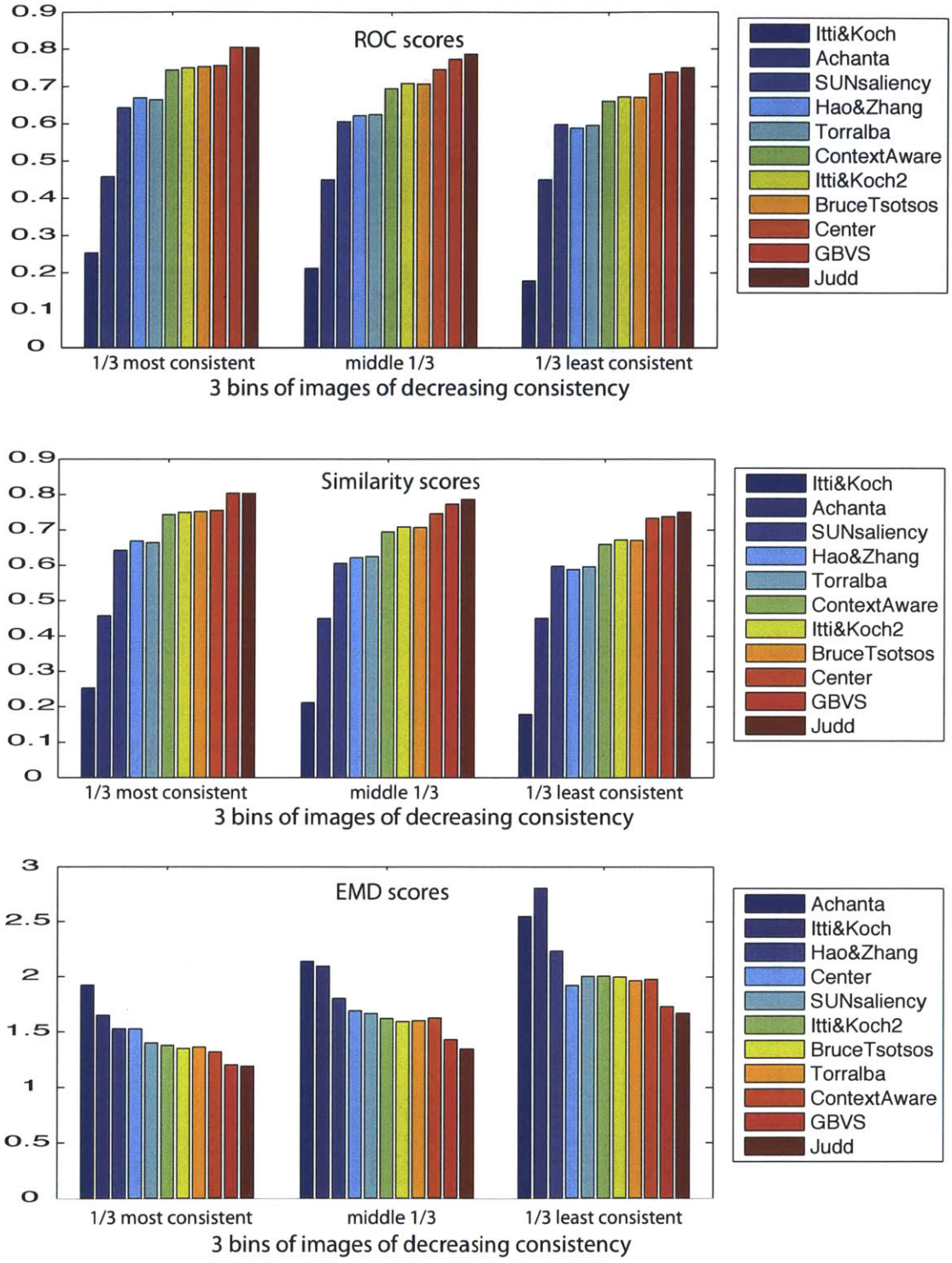

Figure 4-11: The graph plots the performance of the models for all three metrics across different bins of images: most, medium, and least consistent images. As images become less consistent, performance scores get slightly worse.

difference in these two groups is related to whether or not the model includes a notion of the bias of human fixations toward the center.

In addition we measure the similarity of saliency models to each other and find that many models are similar in nature but that the Achanta and Saliency Toolbox implementation of the Itti and Koch model are different and appear as outliers when the models are plotted in multidimensional space.

We rank our images with respect to the consistency of the human fixations on them and find that this dimension has a small effect on the performance of the models to predict fixations: the more consistent the human fixations, the higher the performance of the model to predict fixations.

The best models still do not predict fixations as well as humans predict other humans' fixations, so there is still room for improvements in models of saliency. To aid in benchmarking these new models accurately against existing models, we provide a data set with many observers and an online site to upload and evaluate new models.

We learn a couple things from the results of this work: One is that combining several features together in a model typically gets better predictive results, possibly at the expense of time. We see directly that adding object detectors and a notion of the center bias create high performances (as in Judd and GBVS models). In addition, blurrier saliency maps give better overall results as seen by the higher performance of the blurrier models (GBVS, Itti and Koch2, Bruce and Tsotsos, Context Aware) over the more detailed, higher frequency models (SUN, Torralba, and Achanta). Instead of delineating salient regions from non salient regions in a binary fashion, having a natural decay is preferable. The nearby regions are likely to be salient and should be ranked higher than regions far from very salient objects. In the future we hope to optimize the blur function for each model before comparing them to each other. This would help us understand which model fundamentally chooses the right locations instead of simply preferring models that are blurrier.

We measure models and make conclusions about their performance by their ability to predict where people look. There are two possible issues with this approach.

The first is that we only have 20 observers per image and therefore do not have a perfect measure of where the average viewer will look. As the number of observers per fixation map increases to infinity, the similarity score between them will be 1. As it is impossible to get infinity viewers, a finite amount of viewers is necessary. For accurate reflection of where the average viewer will look, we would ideally have a similarity score between two fixation maps of 0.9. We estimate that attaining this score would require over 100 observers as the similarity score increases linearly with exponential increase in viewers. One future area of research is finding ways to get many more observers on experiments.

The second issue is that there are very different but still very legitimate ways to measure model performance. An alternative is to evaluate their use in applications. If the system performance is increased in either time or quality due to the model, it is not necessarily important to achieve exact correspondences to human eye movements.

Chapter 5

Fixations on low-resolution images

Abstract

When an observer looks at an image, his eyes fixate on a few select points. Fixations from different observers are often consistent–observers tend to look at the same locations. We investigate how image resolution affects fixation locations and consistency across humans through an eye-tracking experiment. We showed 168 natural images and 25 pink noise images at different resolutions to 64 observers. Each image was shown at eight resolutions (height between 4 and 512 pixels) and upsampled to 860 x 1024 pixels for display. The total amount of visual information available ranged from 1/8 to 16 cycles per degree, respectively. We measure how well one observer's fixations predict another observer's fixations on the same image at different resolutions using the area under the receiver operating characteristic (ROC) curves as a metric. We found that: (1) Fixations from lower resolution images can predict fixations on higher resolution images. (2) Human fixations are biased toward the center for all resolutions and this bias is stronger at lower resolutions. (3) Human fixations become more consistent as resolution increases until around 1664 pixels (1/2 to 2 cycles per degree) after which consistency remains relatively constant despite the spread of fixations away from the center. (4) Fixation consistency depends on image complexity.

5.1 Introduction

It is well understood that fixation locations are heavily influenced by both low-level image features and top-down semantic and task-driven factors. However, what is not well known is how the fixations are affected by lowering the resolution of the image.

Some researchers have studied image understanding at low resolution. Bachmann [1991], Harmon and Julesz [1973], Schyns and Oliva [1997], and Sinha *et al.* [2006] have done face perception studies that show that when an image of a face is downsampled to a resolution of 16 x 16 pixels, viewers are still able to identify gender and emotion reliably. Others have shown that we can essentially understand images, or at least the gist of the images [Friedman, 1979] [Oliva, 2005] [Wolfe, 1998], at a very low resolution [Castelhano and Henderson, 2008] [Oliva and Schyns, 2000] [Oliva and Torralba, 2001b] [Potter, 1975] [Potter and Levy, 1969]. Torralba [2009]

showed that viewers can classify the scene of an image and identify several objects in an image robustly even when the image has a spatial resolution as low as 32 x 32 pixels.

If we understand the gist of the scene at a low resolution, our fixations on low-resolution images are likely directed to locations that we expect to see objects of interest [Biederman *et al.*, 1982] [De Graef *et al.*, 1990] [Henderson *et al.*, 1997] [Henderson *et al.*, 1999] [Loftus and Mackworth, 1978] [Neider and Zelinsky, 2006]. Are these fixations likely to land at the same locations as the actual objects of interest in the high-resolution images? We hypothesize that fixation locations should be similar across resolutions and, more interestingly, that fixations on low-resolution images would be similar to and predictive of fixations on high-resolution images.

As further motivation for our work, we noticed that many computational models that aim to predict where people look used features at multiple scales. For the design of future models, it is interesting to get a notion as to whether all levels of image features are equally important.

In addition, these computational models are designed to predict where people look in relatively high-resolution images (above 256 pixels per side) and often are created from and evaluated with fixations on high-resolution images (such as the fixation databases of Judd *et al.* [2009] or Ramanathan *et al.* [2010]. Where people look on low-resolution images is rarely studied, and more generally, how fixation locations are influenced by the resolution of the image is not well understood. In this work, we explicitly study how the resolution of an image influences the locations where people fixate.

In this work, we track observers' eye movements in a free-viewing memory task on images at 8 different resolutions. This allows us to analyze how well fixations of observers on images of different resolutions correlate to each other and sheds light on the way attention is allocated when different amounts of information are available.

5.2 Methods

5.2.1 Images

The images we used for this study were drawn from the image data set of Judd et al. (2009). We collected 168 natural images cropped to the size of 860 x 1024 pixels. As a control, we also created 25 fractal (pink) noise images, with a power spectral density of the form $(\frac{1}{f})^{(5-2*fractalDim)}$, where our $fractalDim$ was set to 1. We chose $fractalDim = 1$ because it most closely resembles the frequency of natural images [Kayser *et al.*, 2006]. For each of the natural and noisy images, we generated eight low-resolution images with 4, 8, 16, 32, 64, 128, 256, and 512 pixels along the height (see Figure 5-1). To reduce the resolution of each image, we used the same method as Torralba (2009): we applied a low-pass binomial filter to each color channel (with kernel [1 4 6 4 1]), and then downsampled the filtered image by a factor of 2. Each pixel was quantized to 8 bits for each color channel. By low-pass filtering the images, we found that the range of colors was reduced and regressed toward the mean. Since

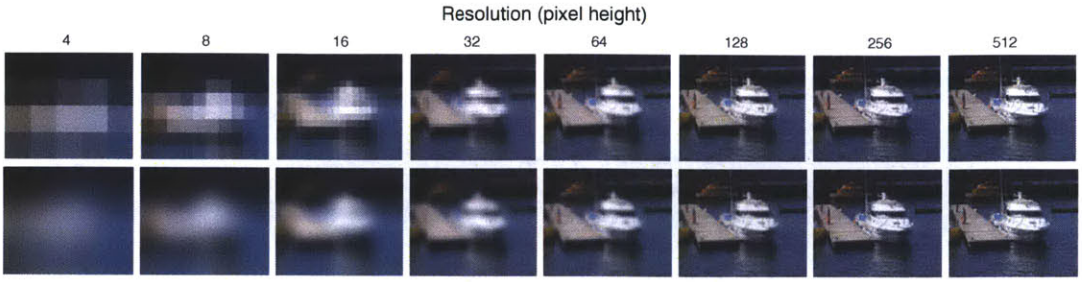

Figure 5-1: Creating the low resolution images. The two rows of images illustrate the amount of information available at each resolution. The top row shows the downsamples images at each resolution (from 4xN to 512xN), and the second row shows the images upsampled to the original size 860x1024. The upsamples images were shown to the participants.

color is an important image feature, we wanted to maintain the range of colors across the blurred versions of a particular image. To do this, we scaled the range of each downsampled image as large as possible within the 0–1 range while maintaining the same mean luminance pixel value. For visualization, the low-resolution images were upsampled using the binomial filter to the original image size of 860 x 1024 pixels. We used code from Simoncelli's [2011] Steerable Pyramid Toolbox to downsample and upsample the images. In total, we had 1544 images (193 images at 8 resolutions). In this paper, we use the size of the downsampled image as a measure of the amount of visual information that is available in the blurred images.

In addition, we separated the natural images into easy/medium/hard bins based on their complexity using the following informal criterion: each image was displayed at several resolutions and the author estimated the lowest resolution at which the image could be understood. The images were ranked such that images understood at low resolution were ranked first, and images understood at higher resolutions were ranked last. The ranked list was then binned into three groups of easy, medium, and hard images. Easy images tended to contain one large object or simple landscape and could be understood at 1632 pixels of resolution. Medium images had multiple objects or more complexity and were understood around 3264 pixels of resolution. Hard images had lots of small details or were often abstract and required 64128 pixels of resolution to understand. Figure 5-2 shows a sample of the natural images in the easy, medium, and hard categories and some noise images, all of which we used in our experiment.

5.2.2 Participants

Sixty-four observers (35 males, 29 females, age range 1855) participated in our eye-tracking study. Each reported normal or corrected-to-normal vision. They all signed a consent form and were paid $15 for their time. Each observer saw a 193-image subset of the 1544 images and never saw the same image at different resolutions. We

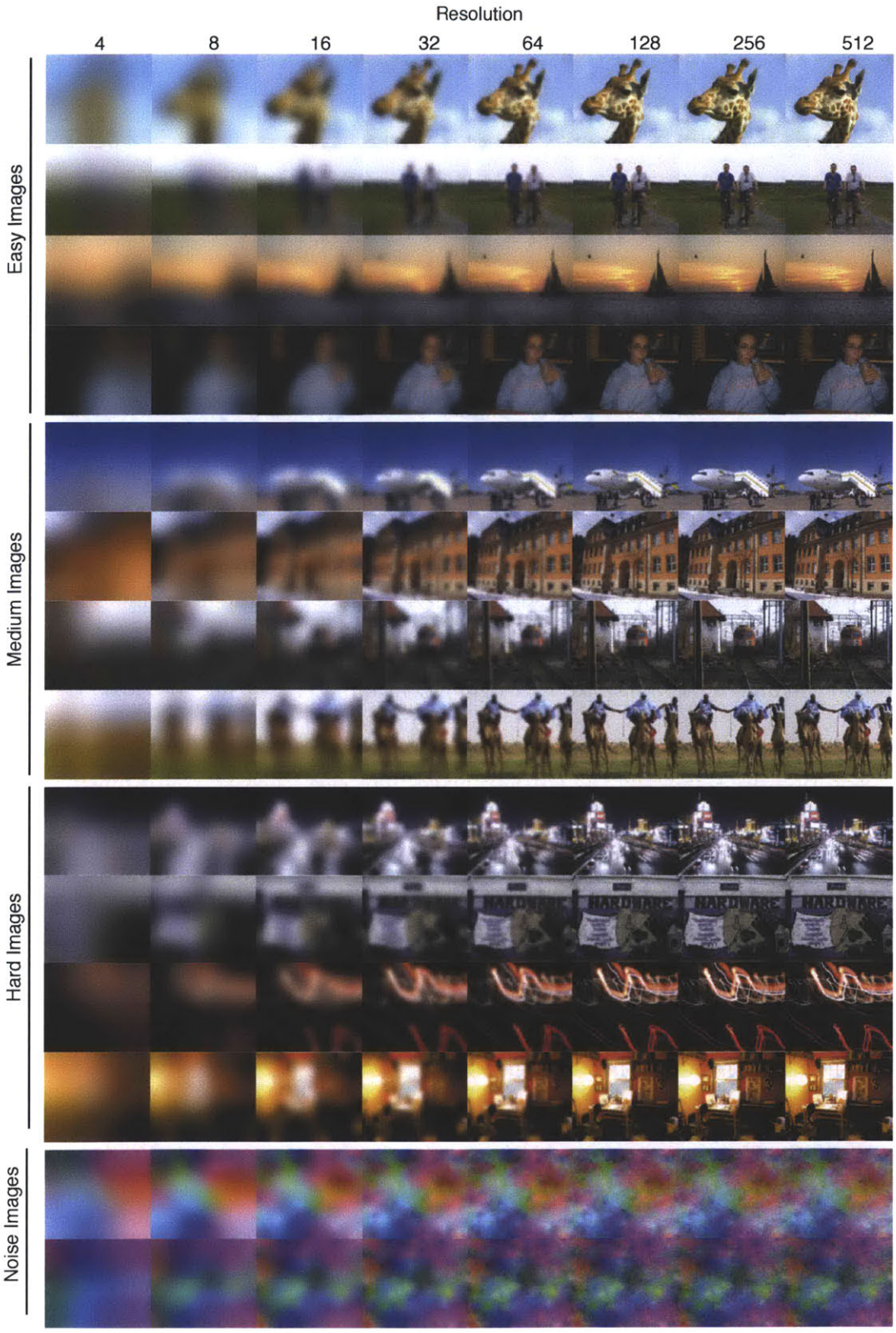

Figure 5-2: Examples of easy, medium, hard and noisy images used in the eye tracking experiment. Note how more resolution is needed to understand the hard images as compared to the easy images. In addition, hard images at high resolution offer more things to look at.

distributed the images such that exactly 8 observers viewed each of the 1544 images.

5.2.3 Procedure

All the viewers sat approximately 24 inches from a 19-inch computer screen of resolution 1280 x 1024 pixels in a dark room and used a chin rest to stabilize their head. A table-mounted, video-based ETL 400 ISCAN eye tracker recorded their gaze path at 240 Hz as they viewed each image for 3 s. We used a five-point calibration system, during which the coordinates of the pupil and corneal reflection were recorded for positions in the center and each corner of the screen. We checked camera calibration every 50 images and recalibrated if necessary. The average calibration error was less than 1 degree of visual angle ($\sim$ 35 pixels). During the experiment, position data were transmitted from the eye-tracking computer to the presentation computer so as to ensure that the observer fixated on a cross in the center of a gray screen for 500 ms prior to the presentation of the next image. We provided a memory test at the end of the viewing session to motivate observers to pay attention: we showed them 12 images and asked them if they had seen them before. This was not used in the data analysis.

The raw data from the eye tracker consisted of time and position values for each data sample. We use the method from Torralba et al. (2006) to define saccades by a combination of velocity and distance criteria. Eye movements smaller than the predetermined criteria were considered drift within a fixation. Individual fixation durations were computed as elapsed time between saccades and the position of each fixation was computed from the average position of each data point within the fixation. The code for identifying saccades and fixations is on our website[1].

We discarded the first fixation from each scan path to avoid the trivial information from the initial fixation in the center. Figure 5-3 shows the fixation locations for eight different observers on some of the images used.

5.3 Results

We have created an interactive web page[2] that allows readers to view the fixation data collected from our experiment and get an intuitive understanding for where people look on images of different resolutions.

Figure 5-4 shows that, as the resolution of the image decreases, observers make significantly fewer fixations. Within 3 s of viewing time, natural images at a resolution of 512 pixels have an average of 7.9 fixations, while images with 4 pixels of resolution have an average just above 5 fixations [paired t-test: $t(167) = 21.2$, $p < 0.001$]. We found that 97% of our natural images have an average of at least 4 fixations. Similar trends hold true for noise images. Having more fixations at high resolutions is understandable since high-resolution images have a lot more details that attract the attention of viewers; there is more to look at. On lower resolution images, people

[1] http://people.csail.mit.edu/tjudd/LowRes/Code/checkFixations.m

[2] http://people.csail.mit.edu/tjudd/LowRes/seeFixations.html

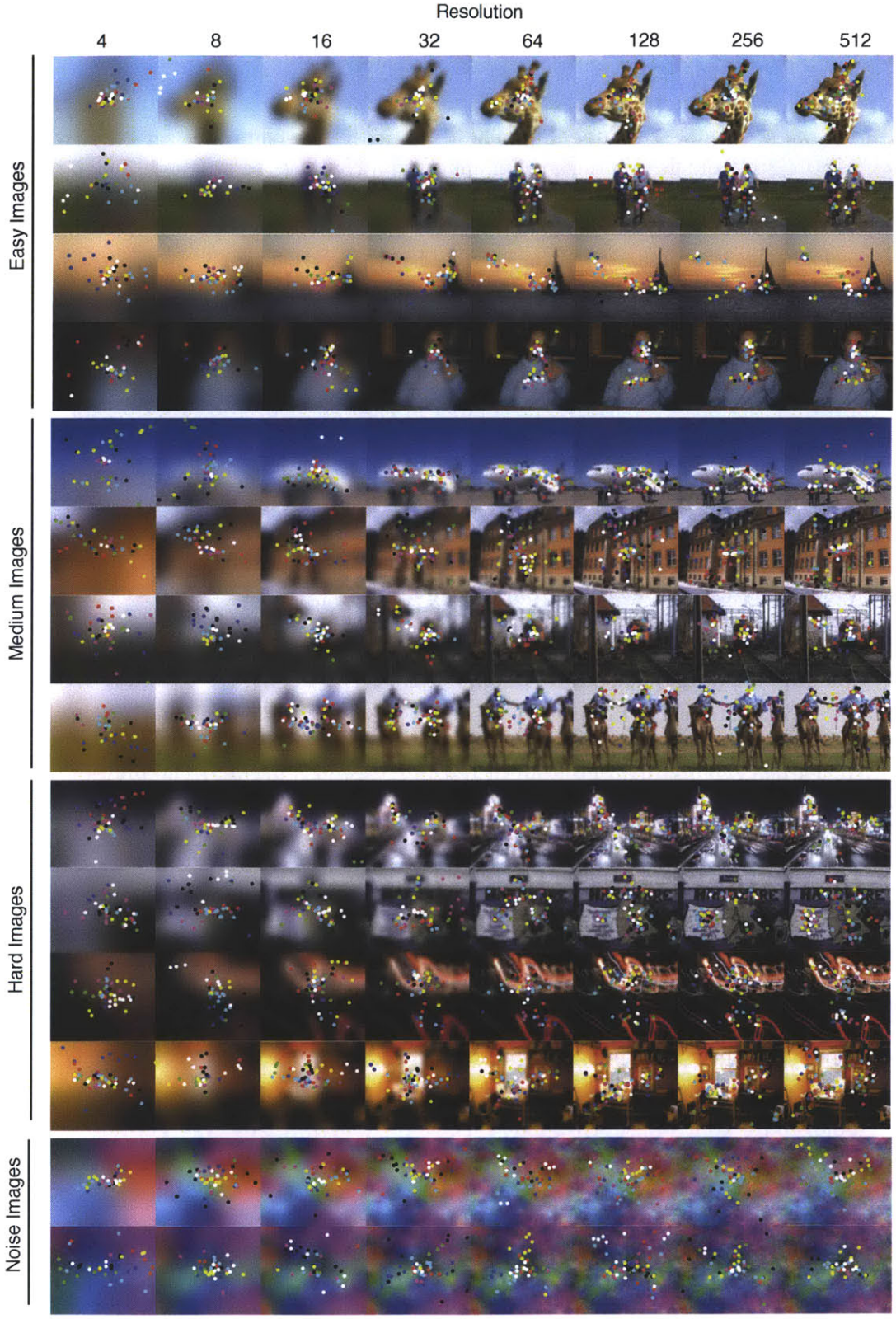

Figure 5-3: Examples of easy, medium, hard and noisy images used in the eye tracking experiment with all fixations from each of the 8 viewers who saw each image. Note that for the easy images, fixations often remain consistently on the primary object of interest even as the resolution increases. On the other hand, fixations spread out to the many details available on the hard images as resolution increases.

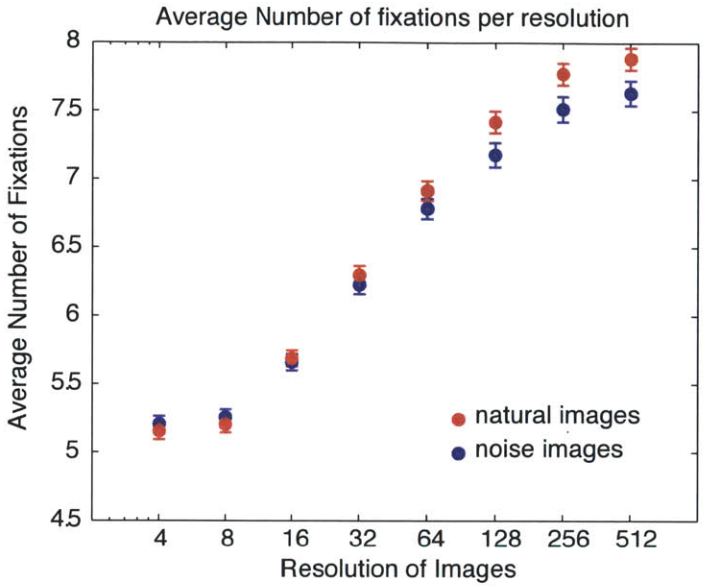

Figure 5-4: The average number of fixations per viewer in 3 seconds of viewing decreases with lower image resolution. The error bars show the standard error over images.

seem to dwell in a location either to try and focus or because they have nothing else to donothing in the image is salient enough to pull their attention away from its current location. We did not observe any obvious effect of the memory test on dwell time.

Figure 5-5 shows a histogram of saccade lengths per resolution on natural images. There is a trend across all resolutions to have many short saccades and fewer long saccades. On average, there are slightly more long saccades on higher resolution images. The plot on the right of Figure 5-5 shows the average saccade length per resolution, with error bars as standard error over subjects. The mean saccade length increases significantly from 4.6 degrees on 16-pixel resolution to 6 degrees on 128-, 256-, and 512-pixel resolutions (paired t-test $t(63) = 10.65$, $p < 0.001$). Interestingly, there is also a small decrease in saccade length as the resolution increases from 5.2 degrees at 4 pixels to 4.6 degrees at 16 pixels (paired t-test $t(63) = 3.63$, $p < 0.001$).

For each image, we create a fixation map similar to the continuous landscape map of [Velichkovsky *et al.*, 1996] by convolving a Gaussian over the fixation locations of each observer who viewed that image (see the fixation map of Image 1 in Figure 5-7). We choose the size of the Gaussian to have a cutoff frequency of 8 cycles per image or about 1 degree of visual angle [Einhäuser *et al.*, 2008b] to match with the area that an observer sees at high focus around the point of fixation. We also made fixation maps with Gaussians of other sizes but found that they did not significantly change the measures of consistency of fixations that we use.

Figure 5-6 shows the average fixation map of all 168 natural and 25 noise images for each resolution. To measure the quantitative difference between the spread of the fixations across the different fixation maps, we measure the entropy of each fixation map intensity image and add that to each fixation map in Figure 5-6. Entropy is

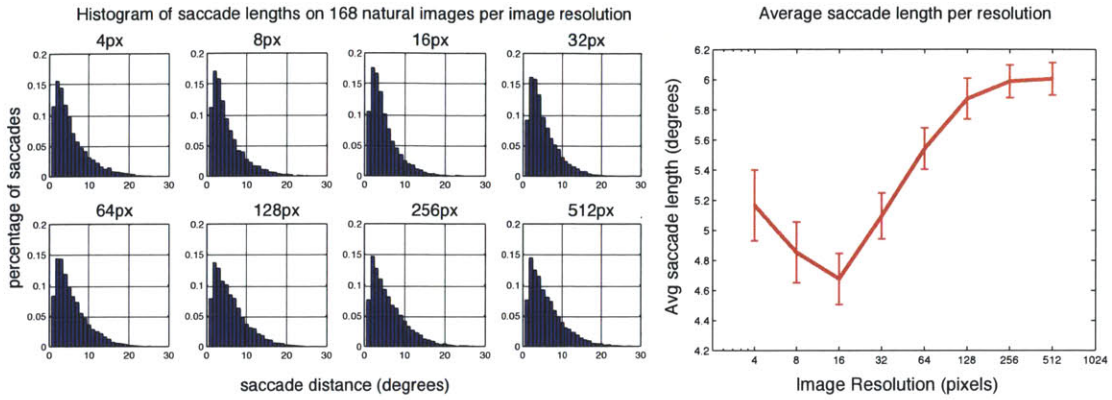

Figure 5-5: Histogram (left) and plot (right) of saccade lengths across resolutions. For every resolution there are more short saccades than long saccades. The average length of saccades increases with resolution. Error bars show standard error over subjects.

a statistical measure of randomness used to characterize the texture of the input image. It is defined as $-\sum(p * \log_2(p))$, where p contains the histogram fractions for the image intensity divided into 256 bins. The higher the entropy, the more spread out the fixations are. In general, researchers have shown that fixations on high-resolution natural images tend to be biased toward the center of the image [Tatler, 2007] [Tatler and Vincent, 2009] [Tseng *et al.*, 2009]. Here in Figure 5-6 we see that as resolution decreases, fixations on natural images get continuously more concentrated at the center of the image. The trend exists for easy, medium, and hard natural image subsets. With noise images, fixations remain consistently biased toward the center of the image for all resolutions.

5.3.1 Measuring consistency of fixations

How much variability is there between observers who look at an image at a given resolution or between observers who look at different resolutions of the same image? To figure this out, we first computed the consistency or agreement among fixations by the 8 separate observers on the same image of a given resolution [Mannan, 1995] [Tatler *et al.*, 2005]. Following the method from Torralba *et al.* [2006], we measured the inter-observer agreement for each image by using the fixations generated by all-except-one observers to create an observer-defined fixation map that was then used to predict fixations of the excluded observer. We use the Receiver Operating Characteristic (ROC) metric to evaluate how well a fixation map predicts fixations from the excluded observer (see Figure 5-7). With this method, the fixation map is treated as a binary classifier on every pixel in the image. The map is thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated. By varying the threshold, the ROC curve is drawn: the horizontal axis is the proportion of the image area not actually fixated selected by the fixation map (false alarm rate), and the vertical axis is the proportion of fixations that fall within

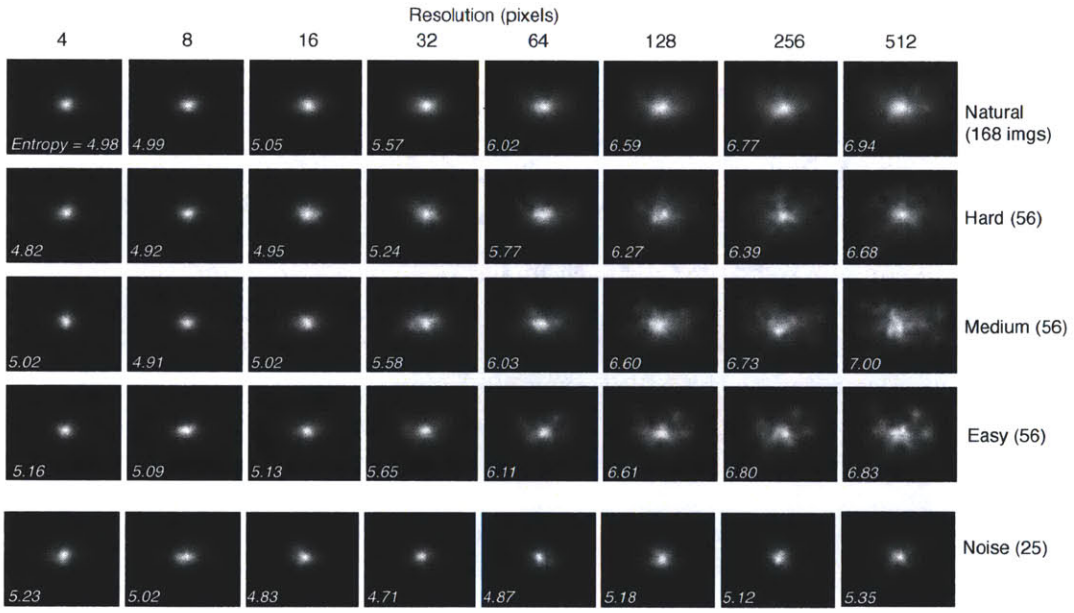

Figure 5-6: Average Fixation Maps. The first row shows the average fixation maps for all 168 natural images for each resolution. In general, as the resolution decreases the fixations become more concentrated at the center. The next three rows show the trend for the easy, medium, and hard subsets of the natural images. The overall trends are the same for each subset. Lastly, the fixation maps for the noise images indicate that fixations are equally biased towards the center independent of the resolution. The entropy of the intensity image for each fixation map is shown in the lower left corner.

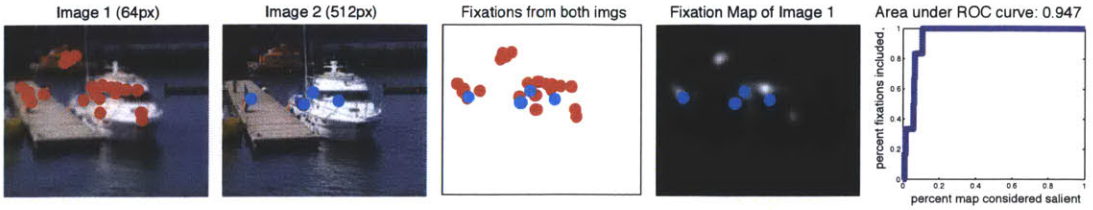

Figure 5-7: Calculating prediction performance. We use an ROC curve to measure how well a fixation map for an image created from the fixations of several users predict the fixations of a different user on the same image at either the same or a different resolution.

the fixation-defined map (detection rate). In this paper, we report the area under the curve (AUC). This one number corresponds to the probability that the fixation map will rank an actual fixation location more highly than a non-fixated location, with a value ranging from 0.5 (chance performance) to 1 (perfect performance) [Harel *et al.*, 2007] [Renninger *et al.*, 2007] [Tatler *et al.*, 2005]. The code we use for calculating the AUC is available on our website[3]. For each image at a given resolution, this process was iterated for all observers. The final measure of consistency among the observers for a particular image at a given resolution was an average of 8 AUC values.

Similarly, we also measure the consistency of observers on different resolutions of the same image. We use the fixations from all-except-one observers on the image at a first resolution to predict the fixations of one of the observers on the image at the second resolution. This is iterated for all 8 sets of 7 observers of the first image predicting each of the 8 observers of the second image, yielding a final measure that is an average of 64 AUC values.

Not all of the agreement between observers is driven by the image–human fixations exhibit regularities that distinguish them from randomly selected image locations. The fixations from all the images in our database are biased toward the center (see Figure 5-6). We can measure how centered the fixations are by using a fixation map of a Gaussian of one cycle per image centered on the image to predict observers' fixations. In this center map or center model, the value of a pixel in the map is relative to the distance of the pixel to the center of the map; pixels at the center are highest and pixels on the edges lowest. We can compare the measure of consistency of different observers' fixations with the performance of the center map to predict fixations.

Using the above methods, we now have a way of computing the consistency of fixations among observers on an image at a given resolution, the consistency of fixations across different resolutions of the image, and the performance of the center map to predict fixations. Since we want to know in general how consistent observers are on each resolution, and how consistent fixations are across resolutions, we create what we call a prediction matrix per image. The rows and columns of the prediction matrix correspond to the varying image resolution from 4 pixels to 512 pixels. Each

[3]http://people.csail.mit.edu/tjudd/LowRes/Code/predictFixations.m

entry in the matrix is the average AUC value indicating how well fixations of the image at a given resolution along the row predict fixations of the image at a given resolution along the column, i.e., how consistent the fixations of the observers are. The diagonal entries show how consistent observers' fixations are on an image at a given resolution (an average of 8 AUC values). The off-diagonal terms measure how consistent observers' fixations are an image across different resolutions (an average of 64 AUC values). As a baseline, we also include the performance of how well the center model and the chance model predict fixations on each resolution. The chance model gives a random value between 0 and 1 for each pixel to create a randomly speckled fixation map.

By averaging together the prediction matrix of all 168 natural images in our database, we get the average prediction matrix in Figure 5-8. The average prediction matrix is computed considering the first 4 fixations of each observer. As an example in reading the prediction matrix, note that, on average, fixations on high-resolution 512-pixel images predict fixations on 512-pixel images with an AUC = 0.92, while fixations on low-resolution 4-pixel images predict fixations on the 512px images noticeably less well with an AUC=0.79.

From this average prediction matrix in Figure 5-8, it is evident that fixations for an image at a specific resolution are best predicted by fixations on the image at the same specific resolution (as seen by the highest average AUC entries along the diagonal). However, it is also evident that fixations on an image can be very well predicted by fixations on images of different resolutions, including lower resolutions. In addition, human fixations are far more consistent than chance, and human fixations are better at predicting fixations than the center model for all resolutions except the very lowest resolutions (4 and 8 pixels).

Because we would like to see how consistency of earlier fixations are different from consistency among all fixations, we show the average prediction maps for a specific number of fixations (1, 2, 3, 4, 6, 8, all fixations) as in Figure 5-9.

5.4 Discussion

When evaluating our data, we start by asking the following two specific questions:

1. How well do fixations on different resolutions predict fixations on high-resolution images? This corresponds to the first column of the prediction matrices of Figures 5-8 and 5-9. We find that fixations on low resolution images can predict fixations on high-resolution images quite well down to a resolution of about 64px. After that performance drops more more substantially but does not drop below the baseline performance of the center map until 16px.

2. How consistent are the fixations across observers on a given resolution? This corresponds to the diagonal of the prediction matrices. We find that consistency varies across resolution. As resolution increases from 4-32px, consistency of fixations between humans increases. After 32px, fixation consistency stays relatively constant despite the spread of fixations away from the center.

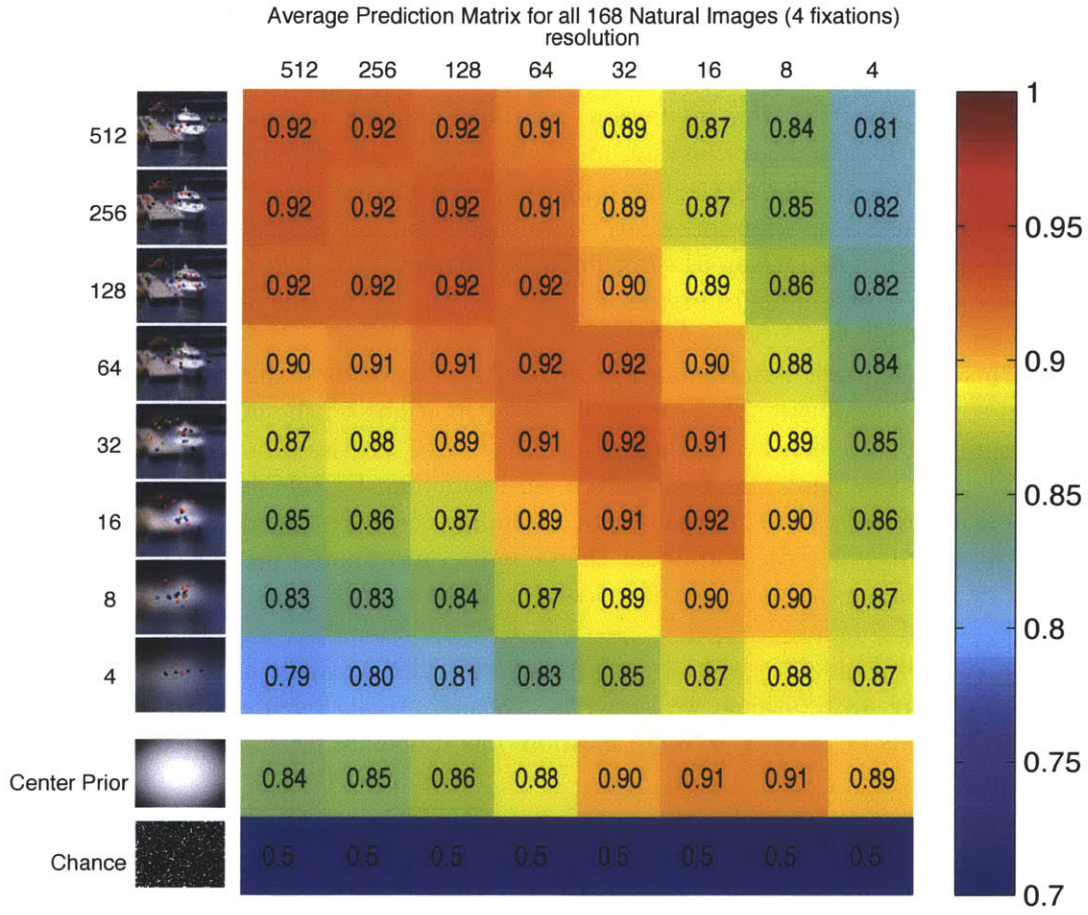

Figure 5-8: Average Prediction Matrices all natural images. This shows how well fixations of images down the rows predict the fixations of the images along the columns. Diagonal entries show how consistent the fixations of the eight observers are and the off-diagonals show how consistent fixations are between observers on different resolutions of the image.

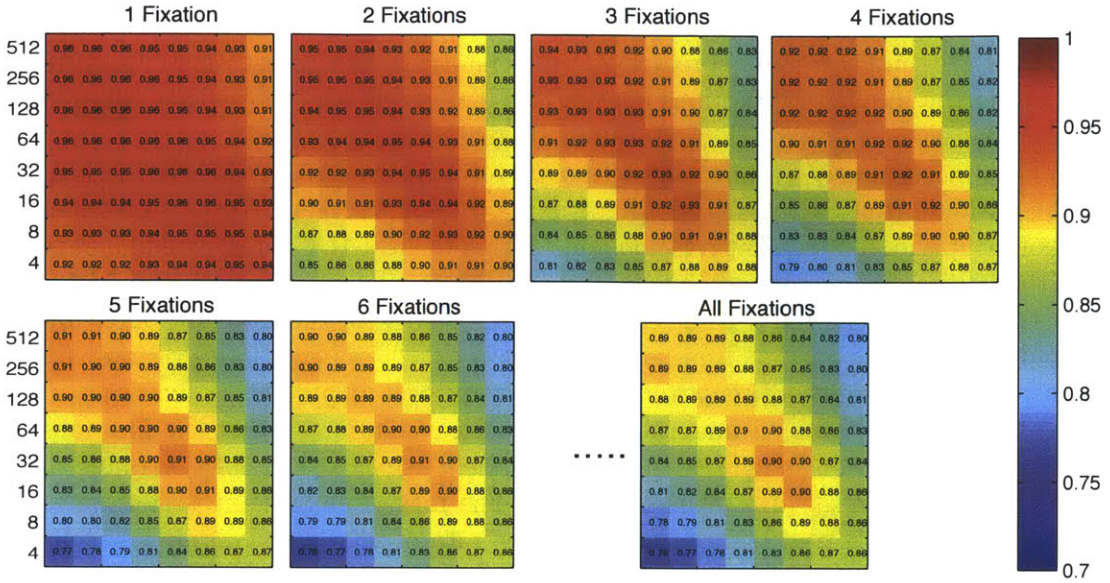

Figure 5-9: Average Prediction Matrices for different numbers of fixations. The first prediction matrix represents how well the first fixations from a given image resolution predict the first fixations on other resolutions. Note that earlier fixations are more consistent than later fixations. Notice also that fixations on the 32 resolution image are most consistent when many fixations are considered.

In addition, we observe the following trends:

1. Fixations across observers and images are biased towards the center of the image and the bias is stronger as the resolution decreases.

2. Human consistency, or the performance of human fixations to predict new human fixations, is almost always higher than several baseline models and artificial models of saliency showing that humans are the best predictors of others fixations.

3. Image complexity affects fixation consistency: the more complex the image, the less consistent the fixations. Image consistency on noise images is very poor.

We explore these results more thoroughly in the following discussion.

5.4.1 Fixations on low-resolution images can predict fixations on high-resolution images

Figure 5-10(a) shows how well fixations from images of each resolution predict fixations on the highest 512-pixel resolution images. The multiple lines represent the AUC values when considering different numbers of fixations per viewer (either 1, 2, 4, or all fixations). Four 1-way repeated measures ANOVAs reveal a main effect of

the resolution [F(7, 441) = 13.8, 35.3, 29.8, 15.4 for 1, 2, 4, all fixations with every $p < 0.001$]. This indicates that prediction performance increases with resolution independent of the number of fixations considered. The graph also shows that the first fixations are easier to predict than all the fixations in absolute terms.

We look more closely at how prediction performance increases with resolution. Note for example, that when considering 4 fixations per observer, fixations on high-resolution 512px images predict fixations on 512px images (AUC=0.92) significantly better than fixations on low resolution 4px images predict fixations on the 512px images (AUC=0.79) [paired t test: t(63)=19.4, p<0.001], or fixations on 32px images predict fixations on 512px images (AUC=0.87) [t(63)=13.4, p<0.001]. However, fixations on lower-resolution images can still do very well at predicting fixations on the 512px image: fixations on images as low as 64px resolution predict the fixations on 512px high-resolution images with an AUC=0.90 (which is 85% of the range of performances AUC=[0.79 (4px)-0.92 (512px)]). The rate of increase in performance is highest between 4px and 64px after which the rate of increase slows down. The average prediction performance of fixations on images of 256px is equal to that of fixations on the 512px image itself.

Figure 5-10(a) also shows that first fixations are more consistent than all fixations. This may be because people tend to look at the most salient locations or objects in an image first [Koch and Ullman, 1985] [Itti and Koch, 2000] [Einhäuser *et al.*, 2008a], and because they tend to look at the center first [Itti and Koch, 2000] [Tatler, 2007] [Tseng *et al.*, 2009] [Tatler and Vincent, 2009]). Earlier fixations are easier to predict in absolute terms, but they are not easier to predict relative to a relevant baseline - the center. There is a larger improvement between the baseline and the human performance for the later fixations.

Our data also shows that fixations on images above 16px predict fixations on 512px images (considering 4, AUC=0.85) significantly better than the center map (AUC=0.84) [t(63)=4.13, p<0.001]. Fixations on lower-resolution images (considering 4 fixations, 8px AUC=0.83) perform significantly worse than the center map (AUC=0.84) [t(63)=7.3, p<0.001]. The fact that humans underperform the center map is due to the number of subjects we consider and the size of the Gaussian map used to create fixation maps. We explore this further in the section on the center map.

Fixations on all resolutions predicted fixations on 512px resolutions better than chance. In the worst case, fixations on 4px image predicted fixations on 512px resolution at AUC=0.79 was significantly better than chance at AUC=0.5.

5.4.2 Consistency of fixations varies with resolution

Figure 5-10(b) shows how well fixations on a given image resolution predict fixations of other observers on the same resolution, i.e. how consistent the fixations are on each resolution. Four one-way repeated measure ANOVAs reveal a main effect of the resolution [F(7, 441)=13.8, 35.3, 29.8, 15.4 for 1, 2, 4, all fixations with every p<0.001]. This result indicates that changes in resolution do affect fixation consistency independent of the number of fixations considered.

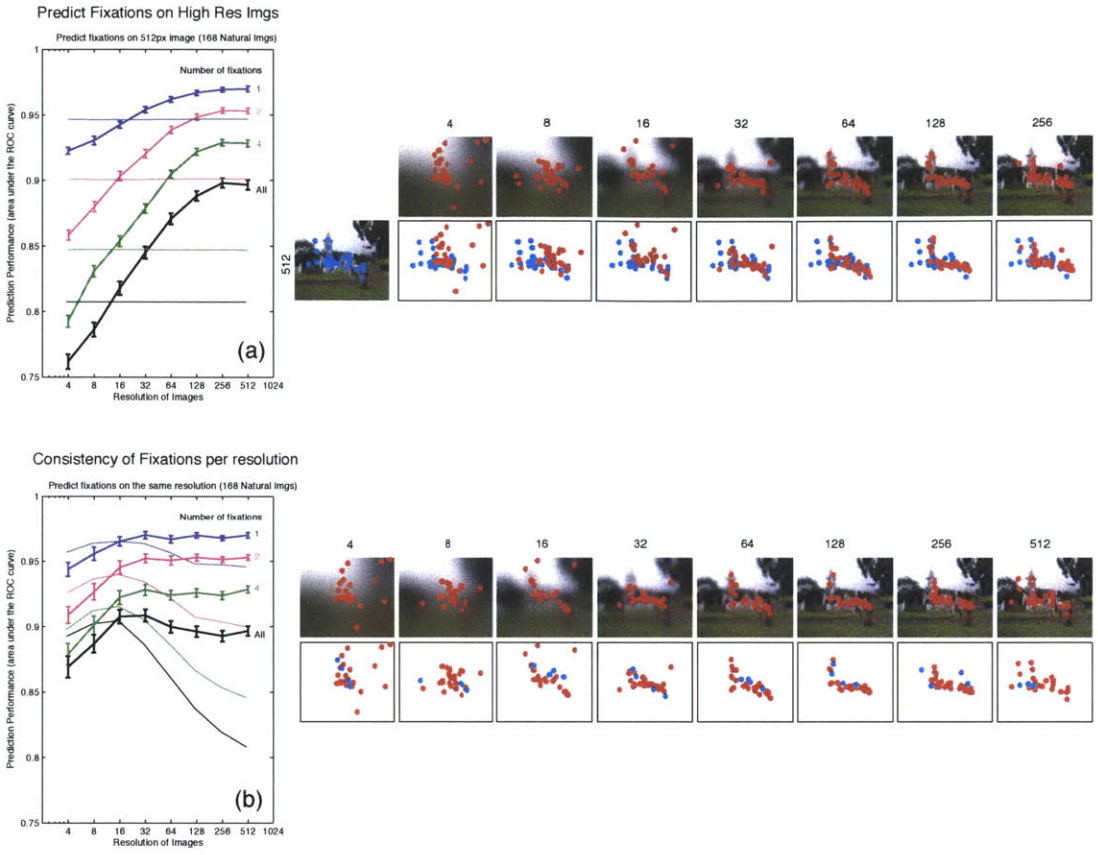

Figure 5-10: (a) The graph on the left shows how well the fixations of each resolution predict the fixations on the high resolution images. This is compared to how well the center map (thin line) predicts high resolution images. In general, fixations on images of 16px and above outperform the center map. Performance increases with resolution, and the rate of improvement slows after 64px; after 64px resolution, you obtain 85% of the range of accuracy available. For visualization purposes, we have shown the fixations from all 8 viewers per image on the right. The ROC performance is calculated as the average of 64 instances of 7 viewers fixations predicting 1 viewers fixations. (b) The graph on the left shows how consistent fixations on images of each resolution are; it shows how well fixations from a given image predict fixations of other viewers on the same image. This is compared to how well the center map (thin line) predicts fixations at each resolution. After 16px of resolution, human fixations outperform the center map. Human performance increases from 4-32px and after that it either plateaus (when considering the first fixations) or declines (when considering all fixations). We believe the plateau happens around the time the image is understood, and the decline happens because viewers' later fixations are spread across the extra details that appear with higher resolution. For visualization purposes, we have shown the fixations of 7 viewers fixations predicting the remaining viewers fixation on the right. The overall ROC performance per resolution for this image is calculated as the average of the 8 of these possible instances.

We investigate more fully how fixation consistency changes with increased resolution. From the graph we see that the average consistency of fixations between viewers increases with resolution from 4-32px (from AUC=0.87 to 0.92 for the 4 fixation line [t(63)=7.5, $p < 0.001$], from AUC=0.86 to 0.90 for the all-fixation line [t(63)=6.5, $p<0.001$)]. Then, for the first 4 fixations, consistency between viewers plateaus after 32px, meaning observers do not get any more consistent with each other with increased resolution. This means that even through the resolution has decreased by a factor of 16 (from 512px to 32px), viewers are as consistent among each other when looking at the low-resolution images as when they look at the high-resolution images. Viewers do not need full resolution images to have consistent fixations. However, when considering all fixations, consistency seems to decrease slightly after 32px from AUC=0.90 to AUC=0.89 at 512px [t(63)=3.2, $p<0.05$].

While the center map outperforms humans consistency for very low image resolutions, humans outperform the center map after 16px. At 32px, observers fixations are significantly more consistent with each other (for the 4-fixation line AUC=0.92) than they are with the center map (AUC=0.90) [t(63)=7.6, $p<0.01$)] and the tendency only grows with resolution. It is interesting to note that while the consistency of the humans plateaus above 32px resolution, it does so as fixations become more spread apart. See from Figure 5-6 that fixations become less centered, and see from Figure 5-10(b) that the center map declines in performance. Despite the fact that fixations spread out, overall consistency remains about constant meaning that the performance of the humans to predict each other increases with respect to the center map baseline. Observers look at salient image features and objects when there is enough resolution to see them, rather than relying just on oculomotor biases.

5.4.3 Performance of the center map is high because fixations are biased to the center

Figure 5-10 shows that performance of the center map is surprisingly high overall. For high resolution images, the center map is way above chance performance of 0.5, and as resolution decreases, the performance of the center map gets even stronger. Why is this? We examine each issue separately.

On high resolution natural images, other researchers have found that the center map produces excellent results in predicting fixations [Zhang *et al.*, 2008] [Meur *et al.*, 2007b] and several previous eye tracking datasets have shown that human fixations have a center bias. To compare the performance of the center map on our dataset with other datasets, we measure the performance of the center map to predict all fixations of all observers on a given image and average over all images in 3 different datasets (see Figure 5-11). Using this method we get an AUC of 0.78 for our 512px natural images, an AUC of 0.803 for the Bruce and Tsotsos [2009] dataset, and an AUC of 0.802 for the Judd *et al.* [2009] dataset. (Note that the AUC=0.78 for our dataset reported here is slightly lower than the AUC=0.8 for all fixations on 512px images reported in Figure 5-10(b) because there we average the performance of the center map per observer.) Ultimately, the center map performs well at predicting

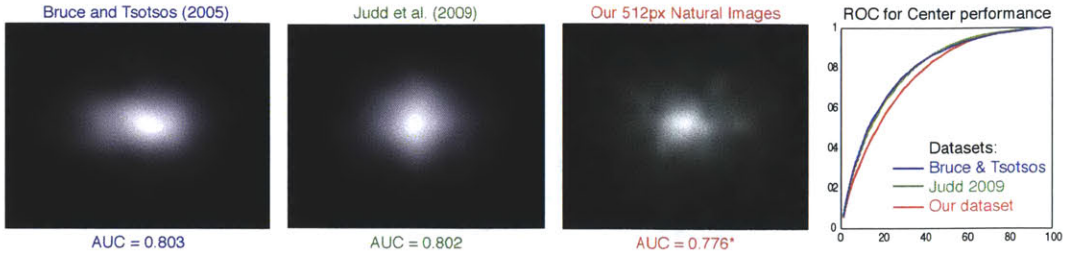

Figure 5-11: The average fixation maps accumulated from all fixations of all users from 3 different image databases show a clear bias towards the center of the image. Because of this strong bias, the performance of the center map to predict human fixations is quite high with an area under the ROC curve around 0.8. *The AUC for our 512px image dataset is 0.78 which differs from 0.8 (center performance of all fixations at 512px) reported in figure 10b because here we calculate center performance to predict all viewers fixations to match calculation for other datasets and there we calculate center performance as the average of the performance for each viewer.

fixations on our database and other databases in the literature.

In Figure 5-12 we compare the performance of humans and the center map with two other related baselines: 1) a fixation map of a randomly selected image of the same resolution, 2) the average fixation map from all images at a given resolution. The graph shows that the average fixation map and the center map have approximately the same performance – understandable given that the average fixation maps approximate center maps as seen in Figure 5-6. The average fixation map actually slightly outperforms the center map because it is slightly rectangular rather than circular and better represents fixations within a rectangular image frame. All three baselines have the same trend: they better predict fixations on low-resolution images than on higher resolution images though their absolute values are different. This indicates that fixations across images are more similar and more centered at lower resolutions than at higher resolutions.

Researchers have studied why fixations are biased towards the center of the image and show that it is due in part to photographic bias (people place objects of interest near the center of their photo composition), viewing strategy (tendency to expect objects of interest in the center), orbital reserve (tendency to look straight ahead), and screen center (tendency to look at the center of a display) [Tatler, 2007] [Tatler and Vincent, 2009] [Tseng *et al.*, 2009]. The photographic bias plays a large role in the center bias of fixations for high-resolution images, but must have a much weaker role on low-resolution images where the principal objects of interest are not recognizable. On low-resolution images, the effects of viewing strategy, orbital reserve and screen center account for the central fixation bias.

The performance of the center map helps us interpret human consistency performance: 1) One of the reasons human consistency is high at 16 and 32px is definitely influenced by the fact that fixation patterns overall are quite similar on these images. 2) Though absolute performance of human consistency remains somewhat constant

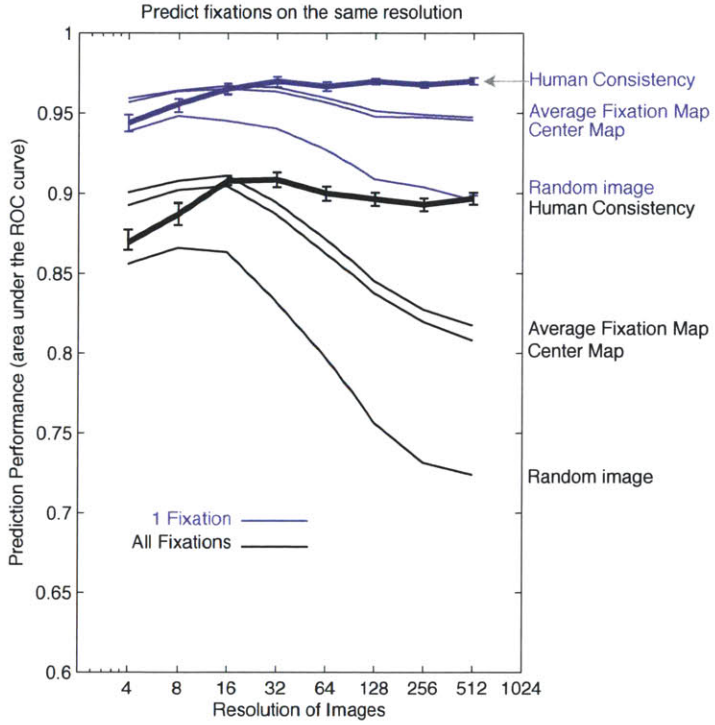

Figure 5-12: Human consistency compared to three baselines. The performance of all three baselines–the center map, average fixation maps and random image maps–have the same trend though their absolute values are different. Performance declines with increased resolution. This indicates that fixations across images are more similar and more centered at lower resolutions than at higher resolutions.

from 32-512px, baselines go down and the relative performance of human consistency over any baseline improves dramatically with resolution (Figure 5-10b). We hypothesize that people are consistent at 32px because fixations across different images look similar (near the center), whereas people are consistent at higher resolution because they look at the same (not necessarily centered) salient locations specific to the image.

One curiosity remains: why is the center map performance higher than human performance on low resolution natural images? Human performance is due to both 1) the number of subjects considered and 2) the size of the Gaussian filter we use to create the fixation map, and these components become more important at low resolution images where fixations are less due to actual salient objects and more due to human biases. The more subjects we consider, the more the human fixation map will match the average fixation or center map and the closer the performance will be to the center map performance. Secondly, the larger the Gaussian filter we consider for creating the fixation map, the more the fixation map matches the center map. When the Gaussian filter is enlarged to 1 to 2 cycles per image, the performance increases for the resolutions 4 and 8 and approaches the performance of the center map for those resolutions. This is reasonable given that the center map is a 1 cycle per image Gaussian located at the center of the image. At 16px and above, the highest performing Gaussian filter was 4 cycles per image, though it was only slightly higher than our original Gaussian of 8 cycles per image. For simplicity, we use 8 cycles per image for all resolutions.

5.4.4 Human consistency and center performance outperform most saliency models

We compare the human performance and center performance with some state of the art saliency models. Figure 5-13 shows that for high resolution images the center model outperforms the Itti and Koch model [2000], Torralba's model [2006] and the Judd model [2009] without the center feature. What is interesting from this analysis is how the performance of the different models change as the resolution of the images increases. In general, saliency models tend to increase in performance while the center decreases in performance as resolution increases. This is not entirely the case for the Judd model without the center feature which rises and then dips with increased resolution. The models that outperform the center model for most resolutions are models that include the center map as a feature in the model. This is the case of the Judd *et al.* [2009] model and the model which combines the Itti and Koch model with the center map. We include the performance of these models for reference, but studying their behavior in depth is beyond the scope of this paper.

5.4.5 Image complexity affects fixation consistency

What causes the plateau or decrease in fixation consistency to appear after 32px of resolution? We discuss two major affects. Firstly, 32px may be the average threshold at which an image is understood [Torralba, 2009]. With less resolution, an image has no semantic meaning and people look randomly at bottom-up salient locations. After

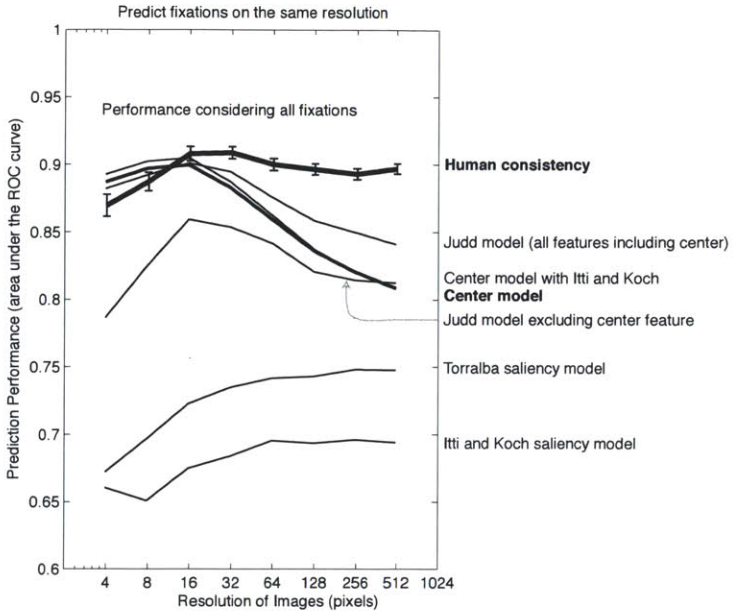

Figure 5-13: Comparison to the performance of saliency models. Human consistency outperforms saliency models in predicting fixations for all models. The center outperforms all models that do not include the distance to the center of the image as a feature of the model.

32px, there is enough low spatial frequency in the image that viewers can understand the gist of the scene [Bar, 2004] [Bar, 2007]. This global information informs about scene category [McCotter *et al.*, 2005] and informs viewers of which objects are likely to be in the scene and where [Loftus and Mackworth, 1978] [Biederman *et al.*, 1982] [De Graef *et al.*, 1990] [Henderson *et al.*, 1999]. It also may be that this is where the primary objects of the scene become visible and thus attract fixations directly [Einhäuser *et al.*, 2008b] [Elazary and Itti, 2008]. One can see this effect in Figure 5-14 where the prediction matrices for specific images have a square-pattern indicating a large jump and then plateau in performance after a given resolution. The resolution at which this jump occurs depends on the content and complexity of the image. Some images are understood early as in 14(a) and others later as in 14(c) and (d).

Secondly, when considering all fixations, 32px has a peak in fixation consistency because it is a threshold at which the image is understandable but not yet full of tiny image details. As image resolution increases, more small details become visible and attract fixations. This is particularly true for images that are more complex in nature. As complexity increases, more resolution is needed to understand what the image contains, and at high resolution there are more things to look at.

Figure 5-15 shows the fixation consistency of all fixations of the three subsets of natural images. A 3x8 (image complexity type x resolution) repeated measures ANOVA on Figure 5-15(b) revealed main effects of the image complexity type [$F(2, 126) = 202.9$, $p < 0.001$], and resolution [$F(7, 441) = 15.4$, $p < 0.001$], and a significant interaction [$F(14, 1512) = 5.6$, $p < 0.001$]. This result indicates that increased

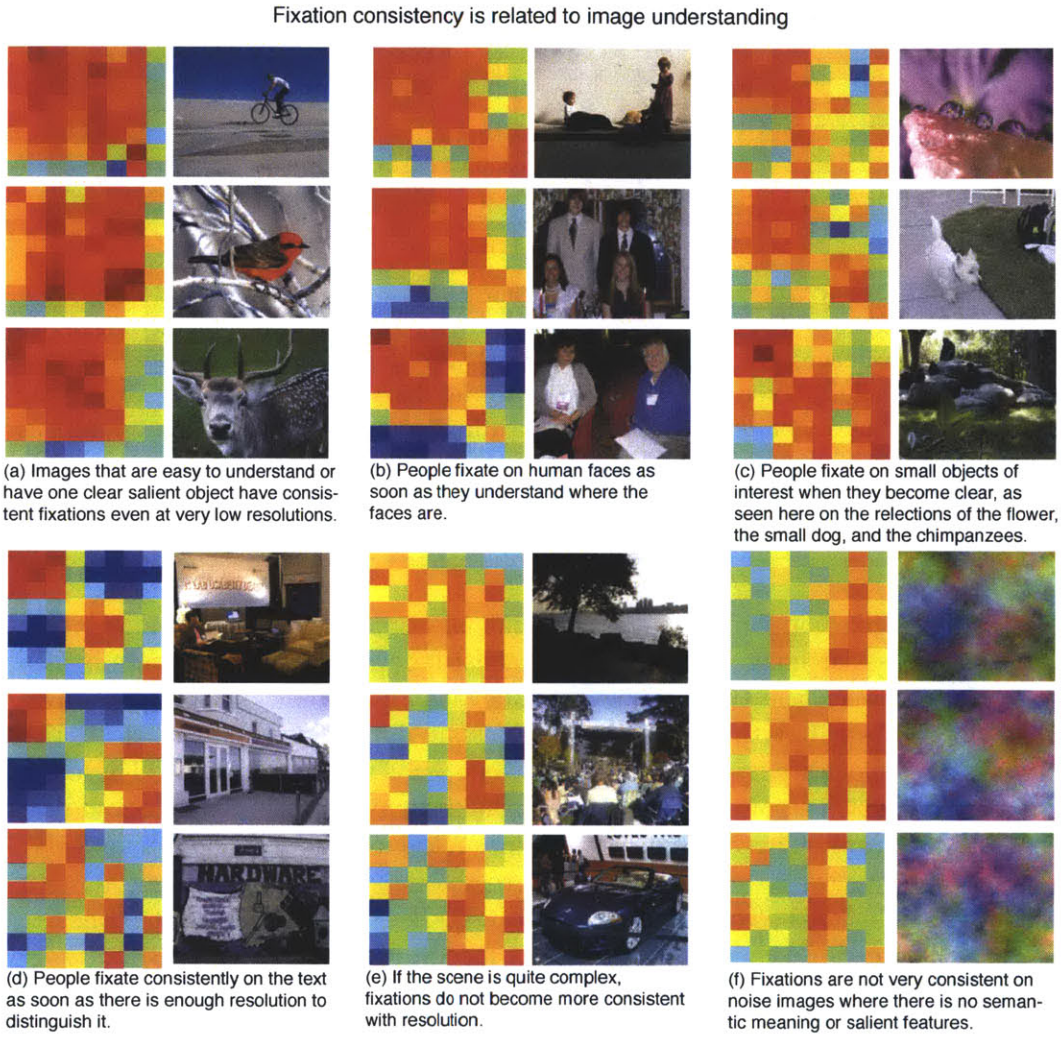

Figure 5-14: These images and their corresponding prediction matrices (for the first 4 fixations) show that fixation consistency is related to image understanding. As soon as there is enough resolution for the image to be understood, people look at the same locations and fixation consistency is higher (the pixel is redder). Different images require more or less resolution to be understood; the more complex the image, the more resolution is needed.

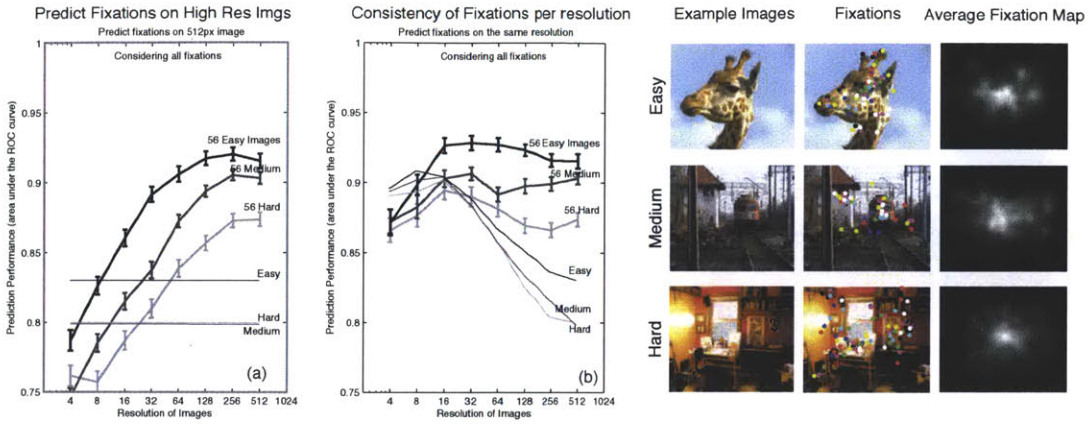

Figure 5-15: Performance on easy, medium and hard images. Here we see the all-fixation trend line separated for different types of images: easy, medium and difficult images to understand. When images are easy, consistency between different resolutions and consistency on a given resolution are much higher than for hard images. In addition, consistency peaks and then declines more strongly for the hard images as resolution increases. In this case, fixations are consistent at low resolution and then become less consistent as the small, complex details of the hard images become visible.

resolution affects fixation consistency, and that the type of image also affects the consistency with easier images having higher consistency. Interestingly, there is also an interaction between the two components showing that predictability goes down as resolution goes up, but this happens faster as the image type gets harder. The decline in fixation consistency appears most strongly for the hard images. These are the images where extra resolution brings extra details which spread fixations out and lead to less consistency. This effect is less evident on the medium and easy complexity images. On easy images, fixations go directly to the main object of interest at both low and high resolution because there is often no other details to look at (consider the image of the head of the giraffe or the girls face in Figures 5-2 and 5-3).

5.4.6 Consistency of fixations on noise images is poor

For comparison, Figure 5-16 (a) (b) show fixation consistency for the pink noise images. We see that for noise images the center map is the best performer: it outpredicts observers on all numbers of fixations. This would change as we add more viewers to the experiment; more viewers would mean that human fixation maps would be closer to the average fixation map, and thus more closely match the predictions of the center map.

Additionally, we see that the curves are mostly flat. This means that predicting fixations on high-resolution images does not increase with fixations from images of increasing resolution; fixations from a low-resolution image are just as good as fixations from a high-resolution image at predicting fixations on high-resolution images.

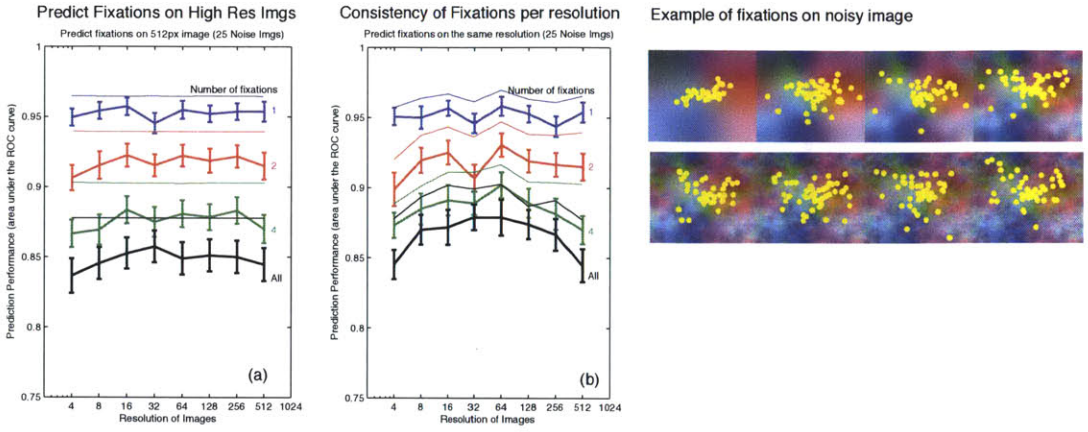

Figure 5-16: Performance on noise images. As a control, we plot the performance of fixations on noise images. Compare these graphs to the performance of natural images. In graph (a) we find that as resolution increases, human fixations do not get more accurate at predicting fixations on high resolution images. In addition, humans never out perform the gaussian center. From graph (b) we see that no particular resolution has more consistent fixations than the others, and once again the gaussian center better predicts fixations than humans for all resolutions.

These noise images, despite matching natural images statistics, have lost natural image semantics and salient features. When these are removed, fixations seem only to be consistent with the center map.

5.4.7 Falloff in visual acuity may affect fixations

It is well understood that the spatial resolution of the human visual system decreases dramatically away from the point of gaze [Anstis, 1998]. The resolution cutoff is reduced at a factor of 2 at 2.5 degrees from the point of fixation, and by a factor of 10 at 20 degrees as seen in Figure 5-17 [Geisler and Perry, 1998]. Since our images extend an angle of 30 degrees, we can approximate the relative visual acuity, and therefore the number of cycles per degree that a viewer can resolve, at different locations in the image.

Given a viewer is fixating at the center of the image, he can resolve 30 cycles per degree (assuming he has normal 20/20 vision) at the center. At 5 degrees of eccentricity, the viewer has 33% of their original visual acuity and can resolve 9 cycles per degree, corresponding to a 256px image. At 15 degrees of visual angle, or at the horizontal edges of the image, the viewer has 14% of visual acuity and can resolve about 4 cycles per degree, corresponding to 128 px.

When a viewer looks at a 512px image, he cannot resolve all the available details in the periphery. When he looks at an image of 64px or below, he is able to resolve all image information even in the periphery because the entire image is below 2 cycles per degree. At this resolution, there is no difference between the center and the periphery

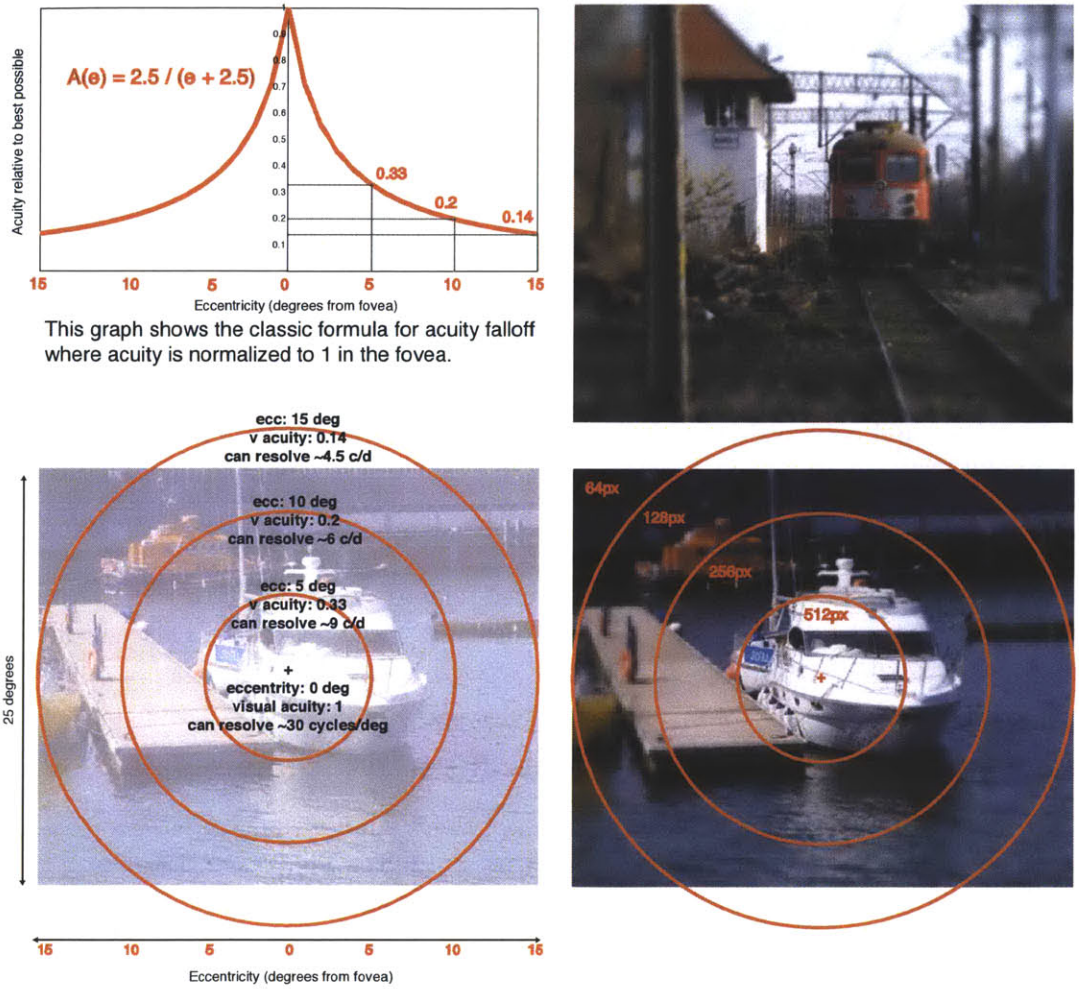

Figure 5-17: Modeling visual acuity falloff. The images on the right simulate the acuity falloff of the human visual system given a fixation at the center of the image. They have high resolution (512px or 16 cycles per deg) at the center and low resolution (64px or 2 cycles per deg) on the edge. They should be seen so that they extend 30 degrees of visual angle. If they are printed on a paper at 3in wide, they should be viewed about 6in away.

in what a viewer can resolve.

We hypothesize that this could have the following implications: 1) For high resolution images, viewers must move their eyes (actively saccade and fixate) in order to resolve different parts of the image. This might not be the case for resolutions 64px and below. 2) The center performs very well for images of 32px and below. In these cases, the resolution is so low (everything is less than 1 cycle per degree) that the entire image can be resolved using peripheral vision; viewers dont need to move their eyes to distinguish details. This could account for the very high performance of the center model on low resolution images. As the cycles per degree get higher, eyes move away from the center in order to resolve the image details.

5.5 Conclusions

We draw the following conclusions from this work:

Firstly, fixations from a specific resolution image are best predicted by fixations on the same resolution image. However, fixations from lower-resolution images can also quite reliably predict fixations on higher-resolution images. Prediction performance drops slowly with decreasing resolution until 64px after which it drops more quickly.

Secondly, fixations between viewers become increasingly consistent on a given resolution as resolution increases from 4-32px. After 32px the consistency plateaus despite the average continued spread of fixations away from the center with increased resolution. We hypothesize that consistency stays strong despite the decreasing center bias of fixations because the image becomes understandable and viewers start to look consistently at the saliency objects or locations.

Thirdly, there is a significant bias for the fixations on all images to be towards the center. For high resolution images, the area under the ROC curve for the center map predicting all fixations is 0.8 and agrees with the performance of other datasets in the literature. As resolution decreases, fixations get more concentrated at the center and the performance of the center map increases. This trend agrees with two other baselines: the performance of randomly shuffled image maps and average fixation maps and indicates that fixations across images are more similar at lower resolutions than at higher resolutions.

Fourth, humans predicting other human fixations outperform any model of saliency that aim to predict where people look. In addition, the center map also outperforms any model that does not include the center as a feature of the model.

Finally, fixation consistency is directly related to the complexity of images. Images which are easier to understand have higher consistency overall and remain high with increasing resolution. Images that are more complex and require more resolution to be understood often have lower overall fixation consistency and decrease in fixation consistency with resolution. We hypothesize that this is because later fixations get diverted to small details.

These trends provide insight into how the human brain allocates attention when regarding images. This experiment shows that viewers are consistent about where they look on low-resolution images and are also looking at locations consistent with

where they look on high-resolution images. These findings suggest that working with fixations on mid-resolution images instead of on high-resolution images could be both perceptually adequate at the same time as being computationally attractive.

This result would be useful for real-time image analysis applications, such as robotic vision, which uses a saliency map to prioritize image locations for further processing. Instead of computing saliency on full-resolution images, the preprocess could be sped up significantly by working instead with low-resolution images.

For future work, it would be interesting to better understand how different computational saliency models predict fixations on varying-resolution images differently and how this depends on whether the saliency model is a bottom-up model or includes top-down image cues. In addition it would be interesting to run an experiment where subjects are both explicitly asked about their understanding of an image and tracked for eye-movements. This could lead to results which more directly support the hypothesis that eye movements are under heavy influence of image understanding.

The work from this chapter was published in [Judd *et al.*, 2011].

Chapter 6

Fixations on variations of images

In addition to our experiments and analysis on fixations on low-resolution images, we also ran experiments aimed at understanding where people look on variations of standard high resolution images. We present the motivating questions, the experiment, a summary of the fixation data, and preliminary results. The raw data is available online and further analysis is left for future work.

6.1 Motivating questions

Does cropping images reduce the bias of fixations to be near the center of the image? Many data sets, including ours, report a bias of human fixations towards the center of the image. This is caused in large part to two factors: the *photographic bias* of photographers to put objects of interest near the center of the image, and *viewing strategy* of viewers to look towards the center as objects of interest are expected to be there and that the optimal place to learn the most about the overall image. To better tease these influences apart, we cropped images such that the center of the crops no longer corresponds to the photographers original frame and therefore reduced the photographic bias of these images. We were curious if fixations would still be biased to certain locations. If there is still a bias towards the center, that would indicate that viewing strategy plays a strong role. We crop original reference images in two ways: by taking 9 equally spaced crops of 2/3rds of the image along a grid formation, and 6 equally spaced crops across the horizontal middle of the image (see Figure 6-2).

When do humans look at the face vs parts of the face? On early eye tracking experiments, we noticed that when a face is shown far away, observers look at the face as a whole with one fixation; when a face is shown close up, observers fixate on the eyes, nose and mouth of the face individually (see Figure 6-1). We were curious about when this transition happens – when do observers change from observing a face with one fixation to needing several fixations? What degree of visual angle does the face need to subtend to force this transition? The broader question this uncovers is how large of an area does a person really "see" with one fixation? It is commonly

Figure 6-1: When faces are shown far away, observers fixate on the face as a whole. When faces are show close up, observers look at eye, nose and mouth with separate fixations. We are interested in knowing what size the face needs to be to change from one situation to the other.

assumed that observers see 1 degree of visual angle in high resolution. How does the size of the face relate to this common threshold? To test this we show images of faces at 5 different sizes and placed faces with equal probability on the left or right side of the image. To see what effect the background had in pulling fixations away from the face, we show the faces both with and without their natural background (see example images in Figure 6-3).

How does the scan path change as a function of the visual angle the picture subtends? When an image is shown large or close up, it subtends many degrees of visual angle. Given that the fovea can only see around one degree of visual angle in high resolution, the observer is required to make many fixations in order to see a large portion of the image in high resolution and get an understanding of the image. When an image is shown small or far away, the image subtends a smaller degree of visual angle. In this case, one degree of high resolution from the fovea extends a larger portion of the image, so more is seen with one fixation. How does this affect how many fixations are made? or where the fixations are made? Are fixations on small images at locations that correspond to fixations on large images or not? To test this, we show images at 6 different sizes (see Figure 6-3). We could not test images at distances farther than the computer screen, but smaller images are a proxy for images farther away as the change in visual angle the image subtends is directly correlated.

Where do people look on images with no content? Where people look is influenced by bottom-up and top-down mechanisms that relate to the image content. In addition, humans have natural biases for where they look. We wanted to extract these natural biases by removing influences of bottom-up and top-down content. We show blank images to observers.

How does resolution affect fixations? Our initial experiments on low-resolution images were part of this experiment. Based on the interesting initial results we ran the much larger experiment and did analysis on fixations on low-resolution images

described in chapter 5. We do not discuss the results on those images here.

6.2 Experimental Setup

We ran the eye tracking experiment under the same conditions as our previously described experiments. We had 30 observers. Each was shown a subset of the cropped, face, low-resolution and small images but never saw a reference image under two different conditions. After the 30 tests, at least 3 observers had viewed each condition of each image. Figure 6-3 shows an overview of the images used in the experiment. This was a free viewing experiment with the promise of a memory test at the end to encourage observers to be alert.

6.3 Preliminary results

6.3.1 Initial results on cropped images

We were curious whether fixations from cropped images would be less centered than that of original reference images. This is possible because the objects of interest from the reference image is less likely to be in the center of the cropped images. We find there is a bias towards the center of the image and towards the center of the reference image, wherever that might be with respect to the crop. In Figures 6-4 and 6-5 we show the average fixation map from each of the different crops. For example, the top left fixation map in figure 6-4 shows the average fixation maps for all images that were created from top left crops of reference images. This fixation map has a bias towards the center and the bottom right of the image. A similar trend happens with the crops along the horizontal axis of the image. The crop on the far left of the reference image has a fixation map with a center bias and a bias that leans to the right.

The dual bias towards the center of the crop and the center of the reference image indicate one of two things: 1) It could be that for a given image, both the photographic bias and the viewing strategy come into play: first the person looks near the center expecting to see objects of interest there and then he moves towards the location of the object of interest which is near the center of the reference crop and therefore on the edge of the crop. 2) It could be that for SOME images where there is a strong object of interest near the center of the reference image, that fixations are near the edges of the crops (indicating fixations are biased to the photographed object) and for other images with no strong object of interest, that fixations are near the center of the crops (indicating fixations are biased due to viewing strategy). It is hard to tell which situation is true given that the shown fixation maps are averaged across all images. It would be good to divide crops into two groups: those that come from reference images with strong objects of interest and those from reference images with no strong object of interest.

For the cropped images, we also asked the question: are some objects so salient that they are fixated on independent of their location in a given crop? Looking at this helps uncover what is intrinsically salient in the image independent of whether it

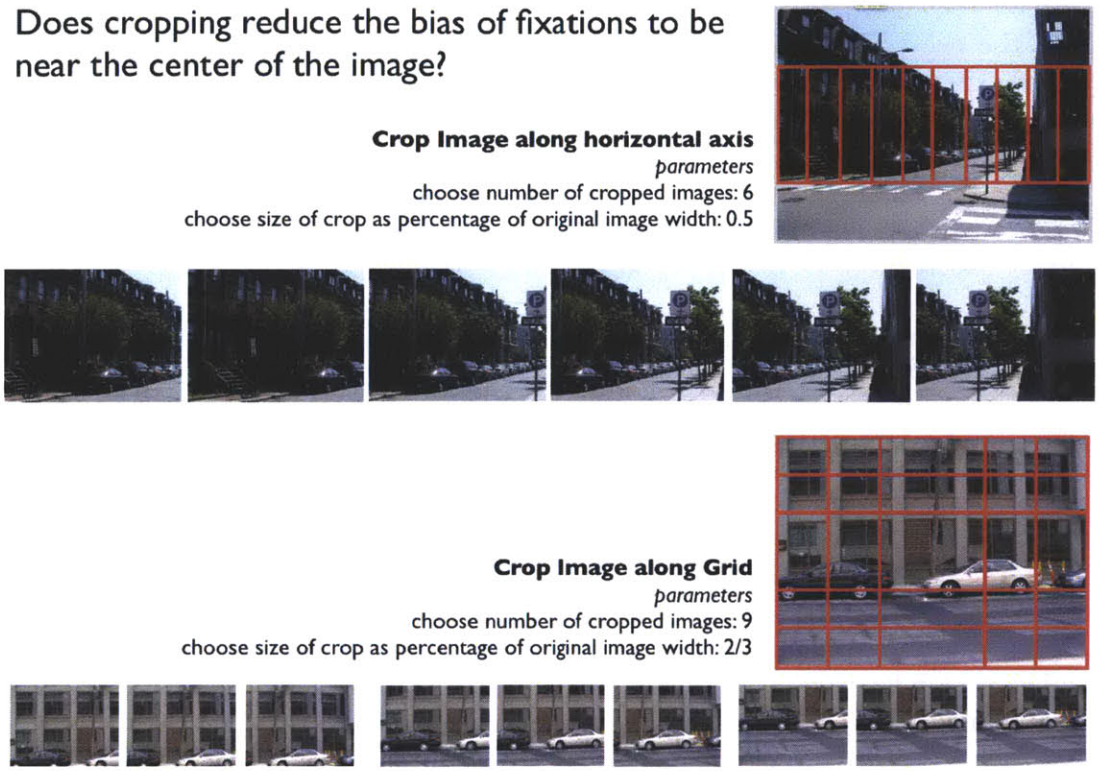

Figure 6-2: These questions motivated our study of variations of reference images.

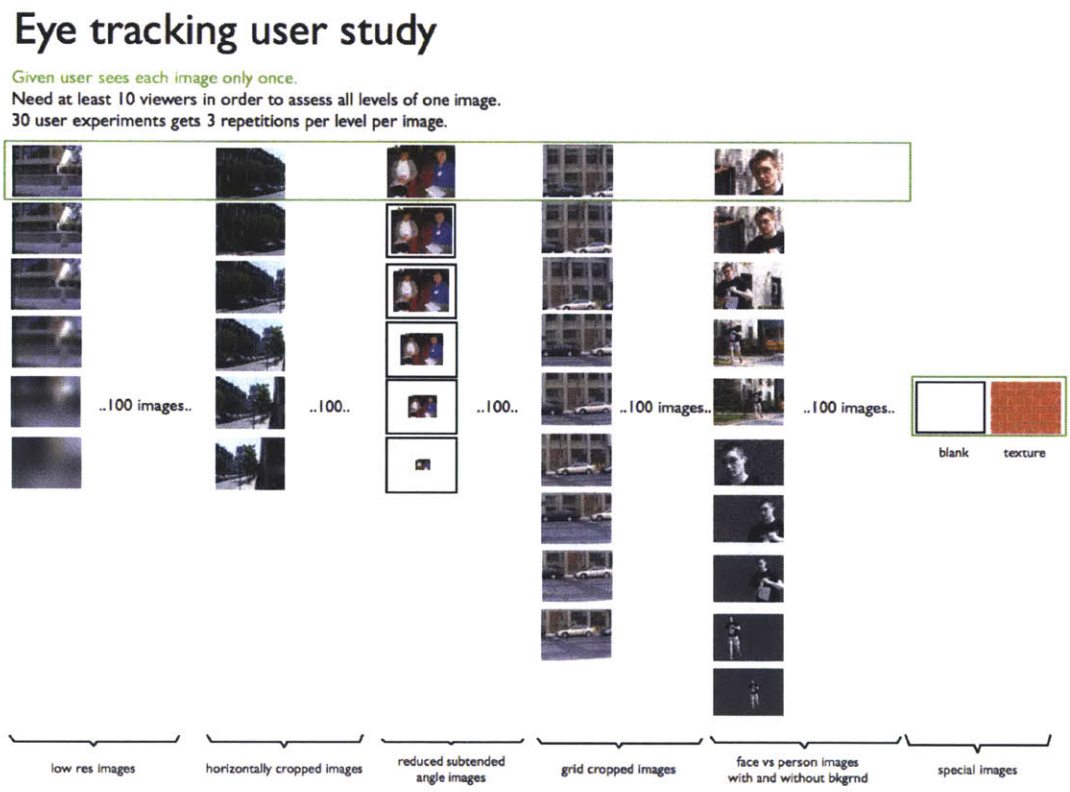

Figure 6-3: We made different variations of reference images to create stimuli used in our experiment. No observer saw two variations of a given reference image. Three observers look at each variation of each image.

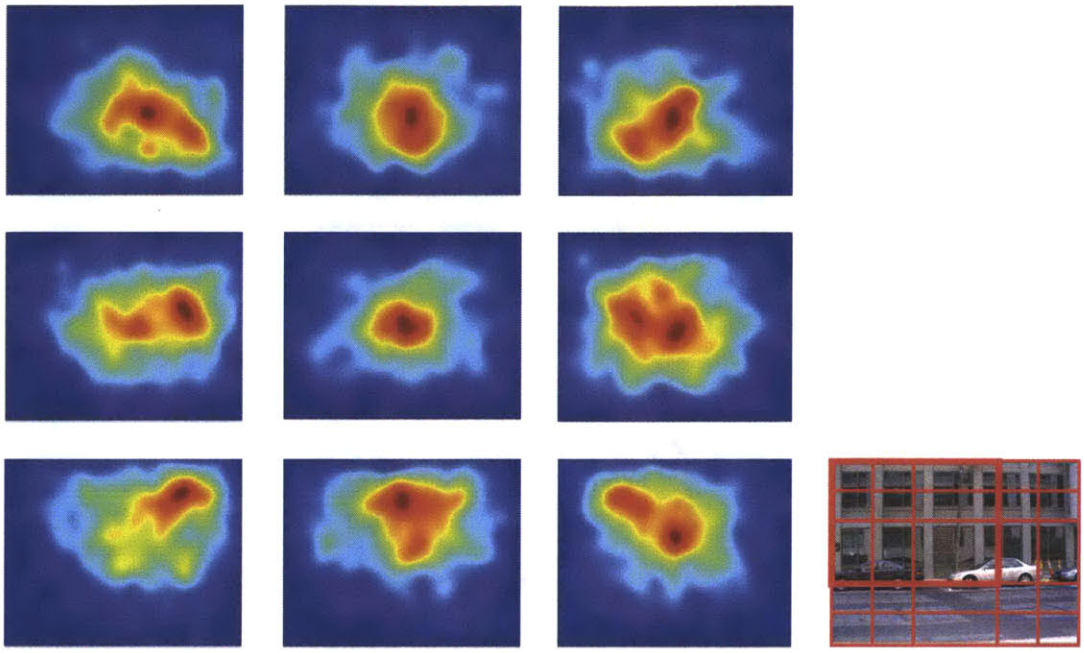

Figure 6-4: Fixation maps from stimuli cropped from reference images. We find a bimodal distribution of the fixations: fixations are biased both towards the center of the crop and the center of the reference image. This indicates that fixations are biased both by viewing strategy to look towards the center, and by the tendency to look at salient objects which in this case are near the center of the reference images.

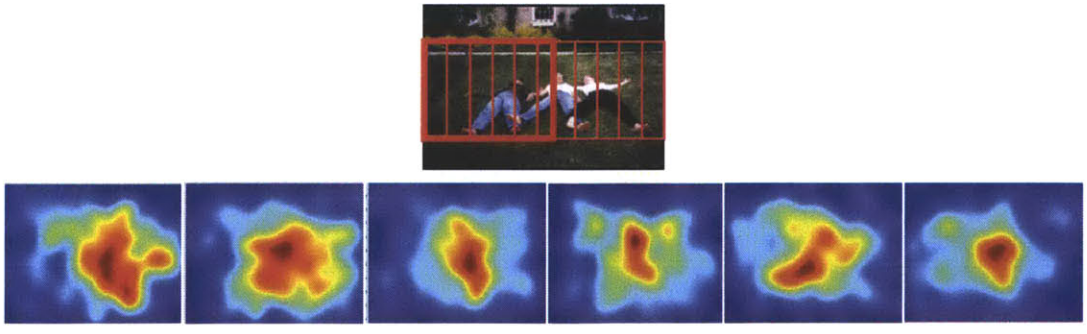

Figure 6-5: Fixation maps from stimuli cropped along the horizontal axis of reference images. Fixations are both biased towards the center of the crop and the center of the reference image.

is located near the center of the image. In figure 6-6, fixations from every crop land on the parking sign and one way sign (the color of the fixation indicates which crop it comes from). Observers looked at the sign even when it was on the far right edge of the image such as in the green crop. A similar effect i seen in figure 6-7 on the right where fixations land on the heads and faces of the camels despite being on the top and sometimes far right or left edges of the crops. These objects are intrinsically salient regardless of their location in an image. On the other hand, when a reference image does not have a particularly salient object the fixations from the crops form a band across the reference image as in the left image of figure 6-7.

Analysis like this helps us discern what objects or locations are fundamentally salient in the image from locations that are fixated on but not particularly salient.

Figure 6-6: Different color fixations indicate which crop they came from. When a given object is fixated on when seen from many different crops, we consider it intrinsically salient. The parking and one way sign are intrinsically salient since observers fixate on them independent of their location in the crop.

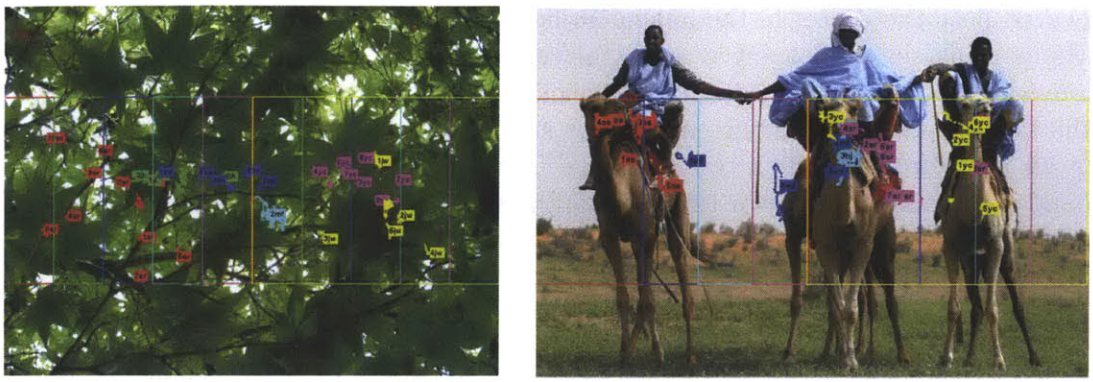

image with band of fixations

image with strong pull to salient locations

Figure 6-7: Different color fixations indicate which crop they came from. Some objects, such are the camel heads are intrinsically salient and attract fixations regardless of their position in crops. Other reference images have no intrinsically salient locations. In that case, observers look near the center of each crop. Fixations make a band across the horizontal axis on the reference image.

6.3.2 Fixations on images of faces

We were interested knowing at what face size would observers change from fixating on individual features to fixating on the face as a whole. From our qualitative assessments observing fixations on face images as in Figure 6-8, we find that people fixate on individual features at the largest face size and when the face subtends half the height of the image. Beyond that, when the face is 1/4 the image height or smaller, observers place only one to two fixations on a face and seem to observe the entire face at once.

As seen in figure 6-9, when the face height is 1/4 the image height, the face subtends about 5 degrees of visual angle. The facial features alone (eyes, nose, mouth) subtend around 2 degrees. Since the fovea sees around 1 degree of visual angle in high resolution, when the facial features subtend 2 degrees of visual angle, the fovea sees most of it in high resolution and may not need to fixate on individual features. Note that our eye tracker is also accurate to about one degree of visual angle, so our measurements at this level are noisy. To get a better assessment of the exact threshold at which observers see all features with one fixation would require further experiments with faces in a subrange between 11 and 2 degrees of visual angle, ideally with a more accurate eye tracker.

We know that observers almost always fixate on a face if there is one in an image. Are faces the only objects people fixate on when they are presented in an environment of other objects? We qualitatively evaluate whether the background attracts fixations away from the face by creating fixation maps of all images with faces at different sizes with and without their natural background, shown in figure 6-10 and 6-11.

We find that when the background is shown (figure 6-10) there are fewer fixations on faces and more on the background, as seen by the spread of the average fixation map to non face areas. When the background is not shown (figure 6-11) there are

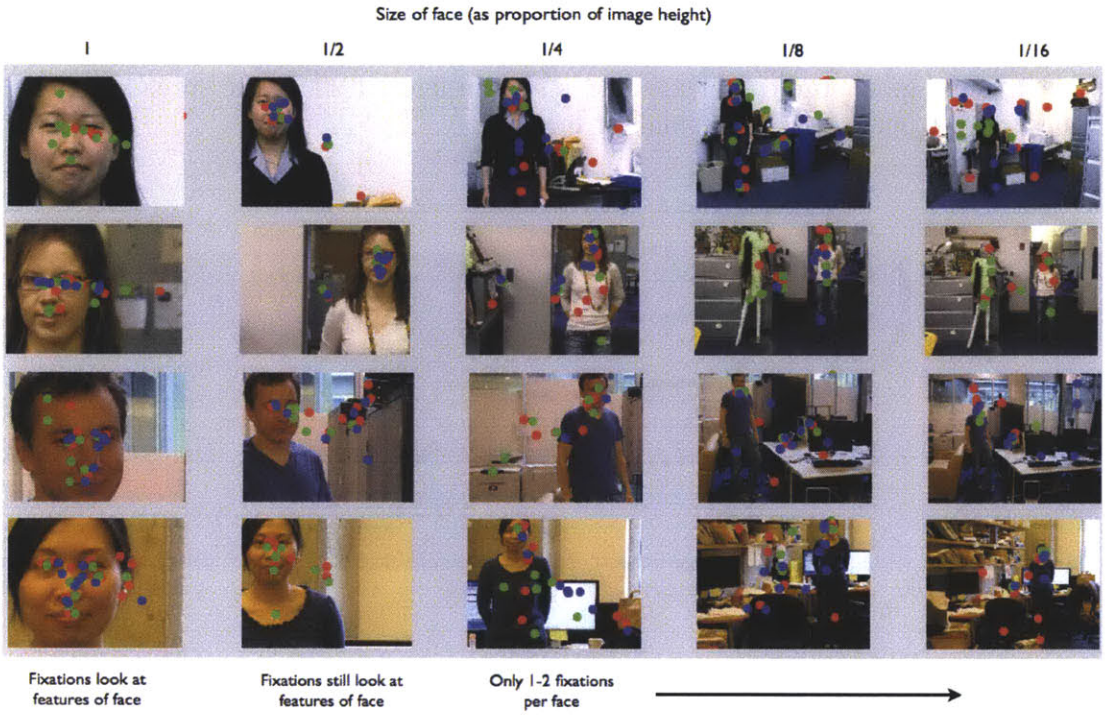

Figure 6-8: Example images with faces at all sizes shown in the experiment, with fixations from three observers each. For images with the two largest face sizes, observers tend to fixate on parts of the face. For images with smaller faces, observers tend to fixate only once or twice on the face indicating that they may see the whole face in a single fixation.

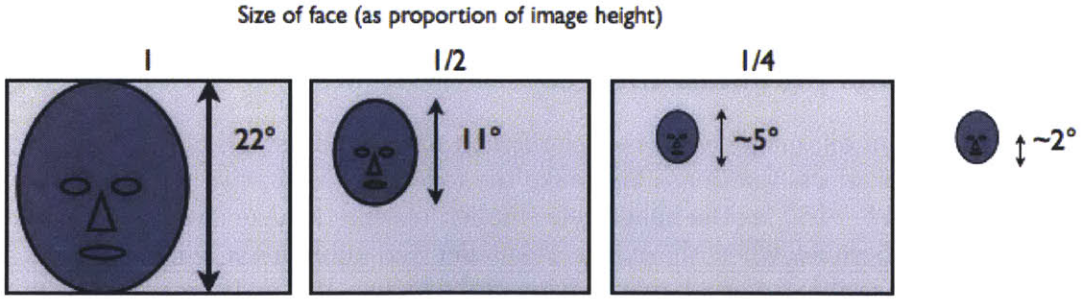

Figure 6-9: The image subtends 22 degrees of visual angle in the vertical direction. The face takes up consecutively smaller portions of the image. When the face is 1/4th the height of the image, the face subtends 5 degrees of visual angle and the facial features alone subtend about 2 degrees.

more fixations on the body. This makes sense as there is nothing else to look at. When the background is shown, we find that people look both at faces and at the background objects. Observers look more at the background as faces get smaller and background gets more complex. This is seen in figure 6-11 as the spread in the fixation map for images with the smallest face size.

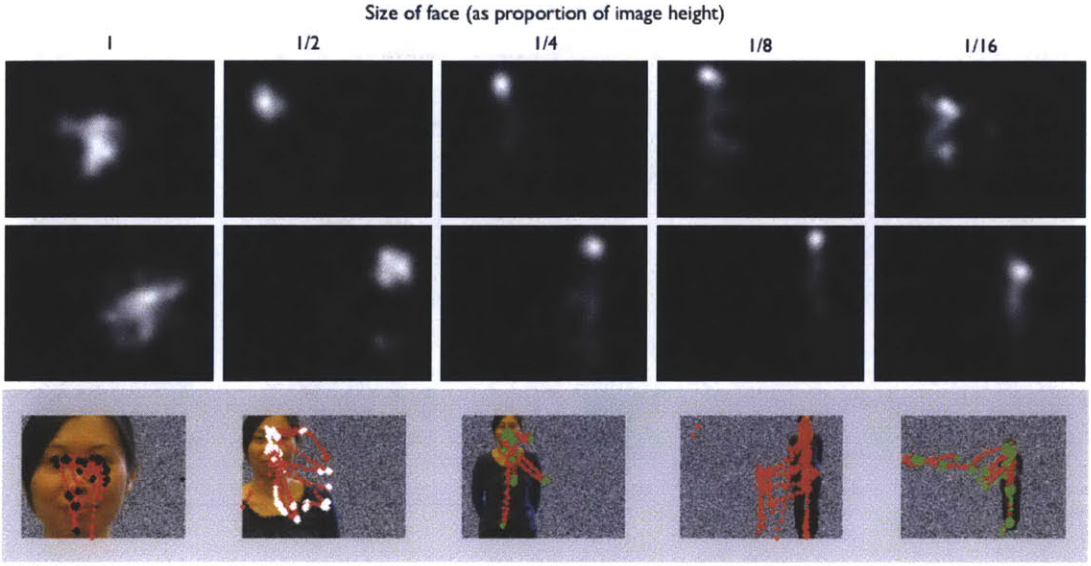

Figure 6-10: The top two rows show fixation maps for images with faces at different sizes shown on a white noise background, separated for faces on the left or right of the image. The bottom row shows one example set of images with fixations from observers. For images with the largest face size, fixations are spread across the face. For all other face sizes, fixations are strongly on faces and somewhat on the body.

6.3.3 Fixations on small images

We measure fixations on images as we reduce them in size to understand how the angle the image subtends affects fixations. We make the following qualitative observations as seen in figure 6-12. First, as the image gets smaller, observers make fewer fixations. This makes sense because when the image is smaller and subtends a smaller visual angle, each fixation is able to see a larger portion of the image. It is no longer necessary to make multiple fixations to view and understand the entire image. Secondly, even though there are fewer fixations, observers look at very similar places as at the larger scale, and specifically they look at the most salient locations in the image. When the image is large, people look both at the salient objects (which are typically larger) and at small details in the background. When the image is smaller, the background details are no longer visible, and only the larger salient objects are visible to attract fixations.

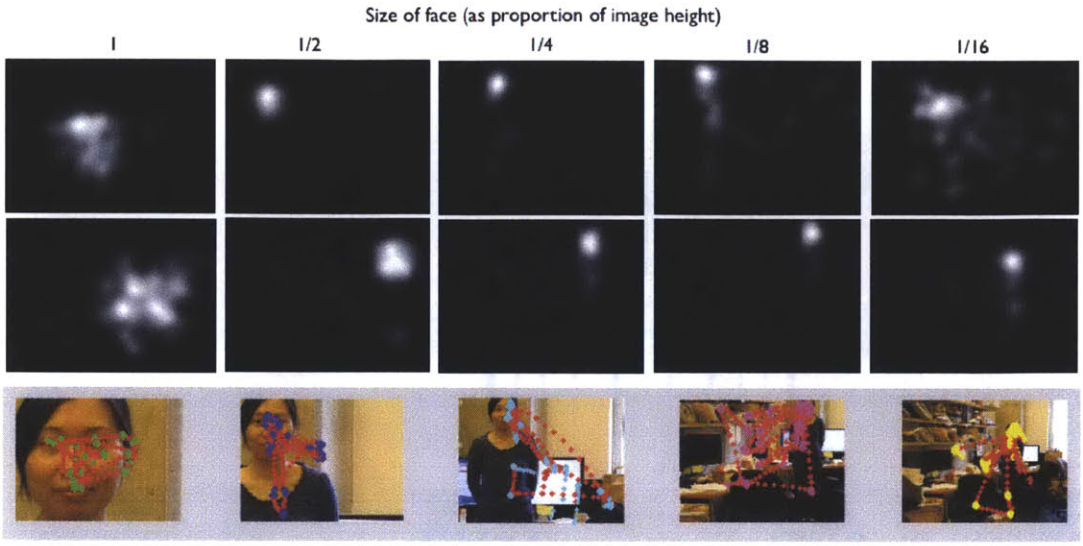

Figure 6-11: The top two rows show fixation maps for images with faces shown with their natural background, separated for faces on the left or right of the image. The bottom row shows example images in this category with fixations. For images with the largest face, fixations are spread on the faces. For images with the 3 mid-sized faces, fixations are mostly on the face with some on the background. Notice they are mostly not on the human body. For images with the smallest face, fixations are much more spread out onto the background.

When the image is very small, our ability to pinpoint the exact location that the observer is fixating on is harder both due to the noise both in the eye tracking measurement and the ability for the observer to see larger portions of the image within the area of one fixation.

Note that the trends on reduced size images match well with the trends on low-resolution images described in chapter 5. When images are small or low-resolution, observers tend to make fewer fixations, and only make fixations on the most salient object. These fixations can match well with fixation locations from larger or higher resolution images. The reason for the similarity in trends is probably because the cycles per degree available at smaller or lower resolution images correlates directly.

It would be interesting to see if these trends can be measured quantitatively. For example, it is possible to create a prediction matrix as in chapter 5 that measures how well the fixations from each image scale predict the fixations at all other scales. In addition, it is possible to see the difference in fixation consistency depending on the complexity of the image. It is probable that simple images with one large salient object would have fixations at more consistent locations across scales than complex images with lots of details.

In figure 6-13 we plot the average fixation map for all images at each scale. There is a center bias of the fixations at all sizes, though this does not get stronger at any particular scale.

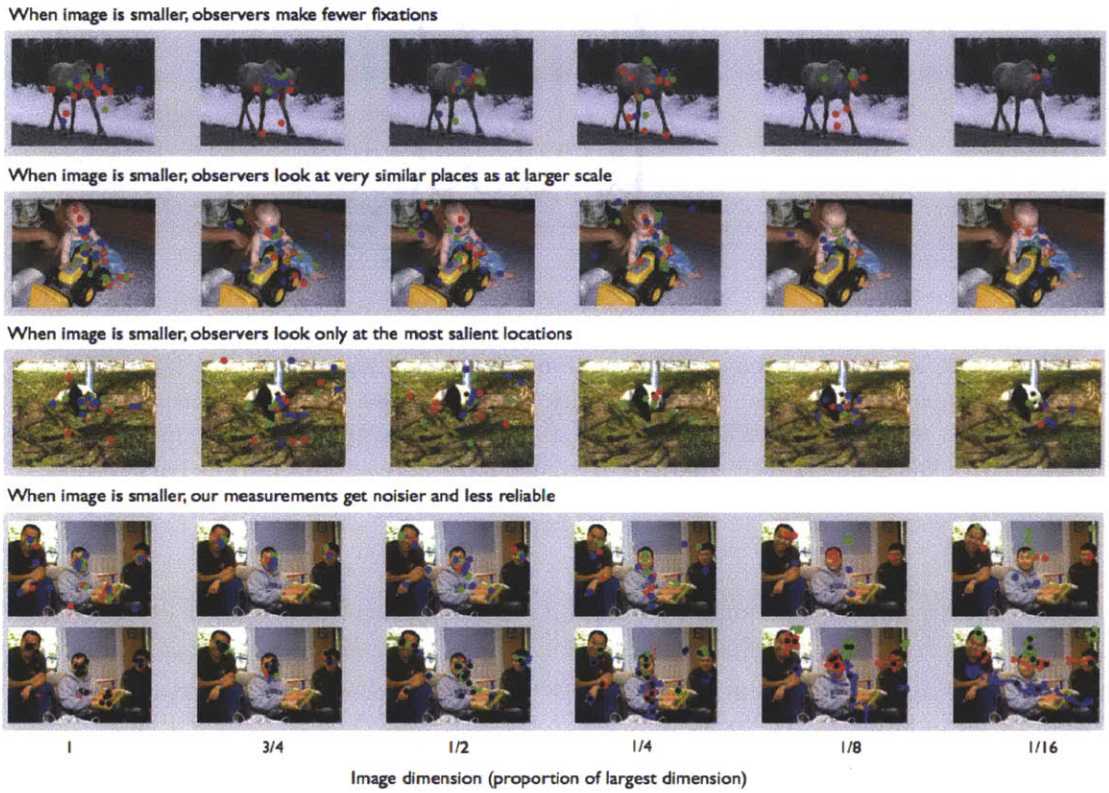

Figure 6-12: Examples of images that were shown at smaller scales in the experiment (here shown all at the same size) with recorded fixations from observers. As images get smaller, observers make fewer fixations, make fixations at similar places as on the larger images, and look at the most salient locations. Note that our fixation location measurements get noisier due to the accuracy constraints of our eye tracker and the ability of the observer to see larger areas with one fixation.

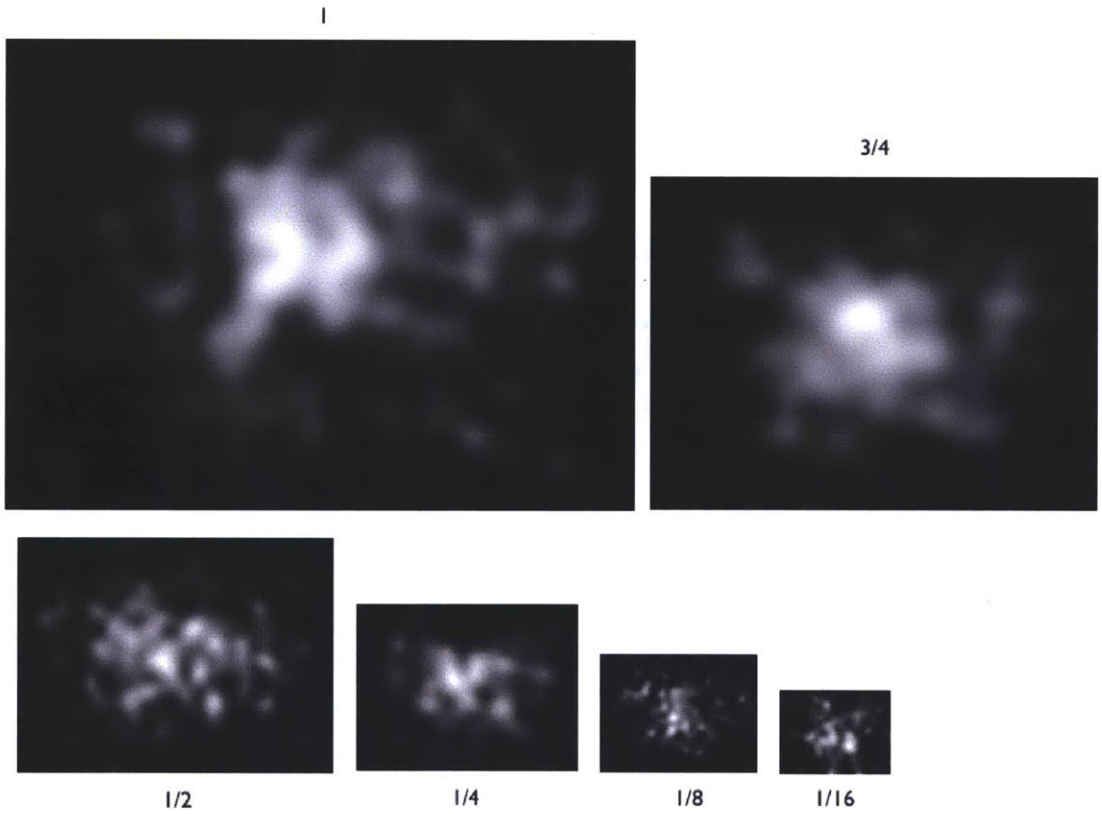

Figure 6-13: The average fixation maps for all images at each scale. The center bias does not seem stronger at any particular scale.

6.3.4 Fixations on blank images

When observers look at blank images, they tend to look at the center of the image, and along the diagonals to the corners of the image. It may be that observers are looking beyond the image and rather at the border between the image and the gray screen the image was shown on. They also tend to look more in the upper half of the image. Beyond the reemergence of the center as a natural bias it is hard to draw deeper conclusions about where people look when there is no semantic content to the image.

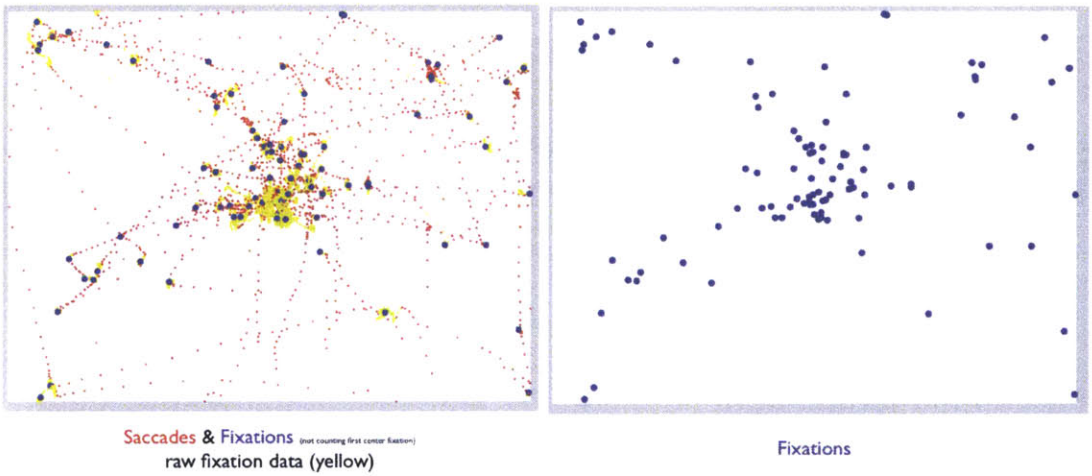

Figure 6-14: Fixations on a blank image. Fixations are biased towards the center, which correlates well with the viewing strategy hypothesis, and to the corners of the image. This may be an artifact of the fact that the white images were shown against a gray background. The difference in color along this edge might have attracted fixations.

6.4 Conclusion

In this preliminary work we look at fixations on different variations of images. Specifically we look at how fixations are affected by cropping, when looking at faces of different scales, and by reducing image size. We found the following exciting initial trends:

1. Cropping does affect fixation location. Instead of fixations only being biased towards the center of the image, fixations tend to be biased towards the center of the reference image. This demonstrates that humans are not incapable of observing the edges of images. Previous data sets have a strong center bias of fixations, but this is due in large part to the photographic bias of salient objects to be near the center rather than the inability for observers to look beyond the center.

2. Viewing strategy does still play a role on cropped images to bias fixations towards the center. Typical fixation maps of cropped images show a bimodal distribution towards one edge and the center. Future work should investigate which affect has a larger role on which types of images.

3. Fixation maps from cropped stimuli can help determine which objects in a reference image are intrinsically salient from those that are randomly fixated on. When an object is fixated on when viewed from multiple crops, it is salient independent of its location in the image. This could be a powerful way of removing affects of the center bias and extracting the most salient locations.

4. When faces subtend about 5 degrees of visual angle, and facial features about 2 degrees of visual angle, observers tend to view all the facial features with one fixation instead of multiple fixations. This aligns well with the assumption that observers see around 1 degree of visual angle in high resolution during a given fixation. Further experiments should be done to refine this estimate.

5. When seeing a face in a natural environment, observers always look at the face, but they do not only look at the face. They also look at the background area. This trend gets stronger the smaller the face is and the more complex the background environment.

6. When images are shown at a reduced scale, and subtend a smaller visual angle, observers tend to make fewer fixations, and place those fixations on the most salient locations. These salient locations align well with where observers look on larger versions of the same images. These trends are similar to trends on low-resolutioin images.

Though it is beyond the scope of this thesis, it would be useful to explore this rich data set further and strengthen the above trends through quantitative evaluation.

Chapter 7

Conclusion

7.1 Summary

Many computational models of attention exist which aim to predict where people look, but they often require tuning many parameters and sometimes do not match actual human fixations. The thesis introduces a new model of saliency that is learned directly from ground truth fixation data. We use a linear support vector machine to learn weights for both bottom-up and top-down features and for a feature that accounts for human biases to look towards the center of an image. We show that by combining these features together we get a model that outperforms all existing models. We note that some features are more important than others: the center feature is most helpful, after which subbands of the steerable pyramid, colors and face detectors make the most improvement to the model.

Research on computational models of visual attention has grown so rapidly in the last 10 years that it is difficult to tell which one works best, as each new model is only compared to a small subset of previous models. Our second contribution is a benchmark which quantitatively measures, using three different metrics, the performance of a broad range of modern models of saliency at predicting where people look. We show that blurrier models, and models that include a notion of the center bias are the highest performing models. The best models still do not predict fixations as well as humans predict other humans' fixations, so there is still room for improvements. To aid in benchmarking new models accurately against existing ones, we provide an online site to upload and evaluate new models.

Most models of attention aim to understand where people look on high resolution images. Our third contribution is an in depth look at how image resolution affects where people look. We find that resolution can be substantially reduced before fixation patterns change dramatically. The work suggests that viewers' fixations become consistent at around the same time as viewers start to understand the image content – and that this can happen at very low resolution. In addition, we show that this threshold depends on the image complexity – images that are easier to understand have high fixation consistency at lower resolutions. This analysis suggests that computational models could work with lower resolution images and still provide similar

performance while gaining speed.

The work of modeling visual attention also requires the use of image data sets with human fixations which can be used to train models and verify performance. This thesis introduces three new data sets of images with eye tracking data. These include the MIT Judd 2009 data set of 1003 natural images with 15 observers, the MIT benchmark data set with 300 images and 20 observers, and the MIT varied-resolution data set with 193 images and 64 viewers spread across different sets of images. All of these are publicly available online[1] and should be helpful to the research community as we move forward in understanding and modeling where people look.

7.2 Open questions

One very important open question is deciding which features are really the most important. Although studied intensively, this is still not fully answered. Most models include some measure of color, orientation, contrast, intensity, which is most important, and what other features should be included? Kootstra *et al.* [2008] found that symmetry is a better predictor for human eye movements than contrast. Baddeley and Tatler [2006] found that characteristics of fixated locations was dominated by high frequency edges. Our work on fixations on low-resolution images in [Judd *et al.*, 2011] and in chapter 5 seems to point out that certain areas are still quite salient despite the lack of high frequency edges (which are blurred at low resolution). In this case context was extremely important: if an observer could make out a person in the image, they would look at the face region even though it was blurred beyond recognition and high frequency detail. Cerf *et al.* [2009] found many fixations land on faces and put a strong weight on the face channel. Zhao and Koch [2011] also found that ideal weights their four feature model depends on the data set being used. For some data sets, the face channels was three times the weight of others, and for other data sets the weight was zero. In our work we found that center map makes the most difference and received the highest weight of all features. Subbands, color and face detection were ranked quite highly as well. Given that a large percentage of fixations in our database landed on text, we hypothesize that text detector would improve performance. Overall there is still much debate and further research is necessary to determine which features are most relevant in which settings.

Another open question relates to how these features are normalized and combined into a final saliency map. Despite the many options available, linear summation of feature channels into the final saliency map remains the norm. Linear summation has some psychophysical support and is simple to apply. However, it may not be the most effective approach. Better models might use higher order functions to combine features. Other construction factors also greatly affect prediction accuracy. Harel *et al.* [2007] note that they include the level of final blur (for instance, the width of a Gaussian blur kernel applied to the master map), and the extent of center bias (for instance, a global center bias multiplicatively applied to the master map) can change

[1] http://people.csail.mit.edu/tjudd/research.html

model performances with up to 30% improvement in some cases. Clearly the way these features are combined into saliency models greatly affects their performance.

Another set of open questions revolves around how top-down cues influence the computation. Top-down influences are not limited to target search, though they are often addressed in this way. Other influences on attention such as the affect of different tasks, or the memories, experiences and internal states of the observers are worth investigating.

Some computational models of attention incorporate face, person and object detectors into the models while other models are used to aid in object detection algorithms. Clearly these two goals are intertwined and it remains an open question to figure out how visual attention and object recognition interact.

Researchers have made much progress by consistently comparing models to human fixations through available data sets. In this thesis we add to this set of data sets. However, it remains an open question whether the available data sets are large enough, good enough and varied enough, and whether or not they have unintentional biases. As in other fields such as object recognition, data sets are useful but often biased [Torralba and Efros, 2011].

The bias of humans to look towards the center of the image is still being explored as well. Do humans really look at the center of images regardless of content or are our databases center biased? Tatler and Vincent [2009], Tseng *et al.* [2009] and Borji *et al.* [2011] suggest the answer is yes on both accounts: our datasets are biased to have objects of interest near the center, but when that is eliminated, observers' natural viewing strategy still leads them to look towards the center. However, this bias could also be a result of the artificial nature of the experiments we run in laboratories and needs to be further explored.

Overall there is still a gap between models and human perfomance. How do we fill that gap? The answer lies in a combination of getting better features, better combining the features, and better understanding the effect of context and of human biases. The way in which we will figure these out and close the gap is either to use a brute force approach of trying all combinations in the spirit of Pinto *et al.* [2009] and by working in concert with neuroscientists to better understand how the human visual system works.

7.3 Future work

There is exciting work to be done extending 2D saliency to 3D scenes, to dynamic scenes, and even more challenging, to interactions of humans in the real world. In addition, to use computational models in the real world, it is also necessary to make them more robust to noise, image transformations, illumination changes, and fast enough to work in real-time. Some work in this area has already happened but there is continued progress to be made.

Another area of future work is optimizing models to be more explicit about overt verses covert attention. Our models currently use the strong eye-mind hypothesis which assumes that what a person is paying attention to is the same as what they

are looking at. While this is a useful first approximation, it is not completely correct: it is possible to pay attention to something you are not looking directly at.

While models of saliency continue to improve, humans remain the best predictor of where people look, so another great area of future research is making eye tracking easier and more accessible. This trend is already in place: In 1967 Buswell's original eye tracker was cumbersome, invasive and not very accurate. Current eye trackers are remote, nonencumbering, non invasive and accurate to within 0.5 to 1.0 degrees. However, because they cost between $5,000-$40,000, they are only used by professionals and researchers. Fortunately there is currently a large effort in making accurate eye tracking devices that use commercial-over-the-shelf cameras and low cost webcams. In fact, in 2006, IPRIZE - A Grand Challenge for Human Computer Interaction, offered one million dollars for a ten-fold improvement in eye-tracking technology while at the same time making it affordable for the average person. Some companies already offer crowd-sourced eye tracking research using webcams at prices that small companies can afford[2].

Having access to easier, cheaper eye tracking will allow for experiments with many more observers. This will allow us to understand what is salient to a large portion of observers to give us a very strong average human fixation map (which would be more reliable than the fixation maps of 20 observers we have now). In addition it helps us fine tune our models to offer saliency maps for more specific subgroups of people: males, females, different cultures, and helps model a larger variety of top down mechanisms.

Some final questions that have come up during the course of exploration for this thesis include: Is there a difference in where people look based on culture or gender or age? Do people who speak different languages look at text of their native language before fixating on text of a different language or different alphabet? How do experts and non-experts look at images differently? Do we follow the gaze direction of people in the images to look where they are looking? Studying fixations on low-resolution images was enlightening because it opened up a new dimension to study fixations on. There are many other dimensions that can still be explored: how do people look at images as they get smaller? as the images get less color? as they see the image repeatedly over time, as they get cropped or retargeted? Clearly there are still plenty of interesting questions to investigate.

[2]For example, see http://www.gazehawk.com/

Bibliography

[Achanta *et al.*, 2008] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Ssstrunk. Salient region detection and segmentation. In Antonios Gasteratos, Markus Vincze, and John Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 66–75. Springer Berlin / Heidelberg, 2008.

[Achanta *et al.*, 2009] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1597 –1604, june 2009.

[Anstis, 1998] S Anstis. Picturing peripheral acuity. *Perception*, 27(7):817–825, 1998.

[Avidan and Shamir, 2007] Shai Avidan and Ariel Shamir. Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3):10, 2007.

[Avraham and Lindenbaum, 2006] Tamar Avraham and Michael Lindenbaum. Attention-based dynamic visual search using inner-scene similarity: Algorithms and bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:251–264, 2006.

[Avraham and Lindenbaum, 2010] Tamar Avraham and Michael Lindenbaum. Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:693–708, 2010.

[Bachmann, 1991] T. Bachmann. Identification of spatially quantized tachistoscopic images of faces: How many pixels does it take to carry identity? *European Journal of Cognitive Psychology*, 3:87 – 103, 1991.

[Baddeley and Tatler, 2006] Roland J. Baddeley and Benjamin W. Tatler. High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research*, 46(18):2824 – 2833, 2006.

[Bar, 2004] Moshe Bar. Visual objects in context. *Nat Rev Neurosci*, 5(8):617–629, 08 2004.

[Bar, 2007] Moshe Bar. The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, 11(7):280 – 289, 2007.

[Biederman *et al.*, 1982] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143 – 177, 1982.

[Borji *et al.*, 2011] Ali Borji, Dicky Nauli Sihite, and Laurent Itti. Quantifying the relative influence of photographer bias and viewing strategy on scene viewing. In *Proc. Vision Science Society Annual Meeting (VSS11)*, May 2011.

[Brandt and Stark, 1997] Stephan A. Brandt and Lawrence W. Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *J. Cognitive Neuroscience*, 9:27–38, January 1997.

[Bruce and Tsotsos, 2006] Neil Bruce and John Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162. MIT Press, Cambridge, MA, 2006.

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

[Buswell, 1922] G.T. Buswell. *Fundamental reading habits: A study of their development.* University of Chicago Press, Chicago, IL, 1922.

[Butko *et al.*, 2008] N.J. Butko, Lingyun Zhang, G.W. Cottrell, and J.R. Movellan. Visual saliency model for robot cameras. In *Robotics and Automation, 2008. ICRA 2008. IEEE International Conference on*, pages 2398 –2403, may 2008.

[Castelhano and Henderson, 2008] Monica S. Castelhano and John M. Henderson. The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, 34(3):660 – 675, 2008.

[Cerf *et al.*, 2007] Moran Cerf, Jonathan Harel, Wolfgang Einhuser, and Christof Koch. Predicting human gaze using low-level saliency combined with face detection. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *NIPS*. MIT Press, 2007.

[Cerf *et al.*, 2008a] Moran Cerf, E. Paxon Frady, and Christof Koch. Using semantic content as cues for better scanpath prediction. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, ETRA '08, pages 143–146, New York, NY, USA, 2008. ACM.

[Cerf *et al.*, 2008b] Moran Cerf, Jonathan Harel, Alex Huth, Wolfgang Einhaeuser, and Christof Koch. Decoding what people see from where they look: predicting visual stimuli from scanpaths. *International Workshop on Attention and Performance in Computational Vision*, 2008.

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

[Chen *et al.*, 2003] Li-Qun Chen, Xing Xie, Xin Fan, Wei-Ying Ma, Hong-Jiang Zhang, and He-Qin Zhou. A Visual Attention Model for Adapting Images on Small Displays. *Multimedia Systems*, 9(4):353–364, October 2003.

[Choi *et al.*, 1995] Yun S. Choi, Anthony D. Mosley, and Lawrence W. Stark. String editing analysis of human visual search. *Optometry & Vision Science*, 72(7), 1995.

[Clark and Ferrier, 1988] J. J. Clark and N. J. Ferrier. Modal control of an attentive vision system. In *IEEE International Conference on Computer Vision*, 1988.

[De Graef *et al.*, 1990] Peter De Graef, Dominie Christiaens, and Gry d'Ydewalle. Perceptual effects of scene context on object identification. *Psychological Research*, 52:317–329, 1990. 10.1007/BF00868064.

[DeCarlo and Santella, 2002] Doug DeCarlo and Anthony Santella. Stylization and abstraction of photographs. *ACM Transactions on Graphics*, 21(3):769–776, July 2002.

[Desimone and Duncan, 1995] R Desimone and J Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995.

[Draper and Lionelle, 2005] Bruce A. Draper and Albert Lionelle. Evaluation of selective attention under similarity transformations. *Comput. Vis. Image Underst.*, 100:152–171, October 2005.

[Ehinger *et al.*, 2009] Krista Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17:945–978(34), 2009.

[Einhäuser *et al.*, 2008a] Wolfgang Einhäuser, Ueli Rutishauser, and Christof Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2), 2008.

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

[Elazary and Itti, 2008] Lior Elazary and Laurent Itti. Interesting objects are visually salient. *J. Vis.*, 8(3):1–15, 3 2008.

[Elazary and Itti, 2010] Lior Elazary and Laurent Itti. A bayesian model for efficient visual search and recognition. *Vision Research*, 50(14):1338 – 1352, 2010. Visual Search and Selective Attention.

[Felzenszwalb *et al.*, 2008] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[Friedman, 1979] Alinda Friedman. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108(3):316–355, 1979.

[Frintrop *et al.*, 2007] Simone Frintrop, Maria Klodt, and Erich Rome. A real-time visual attention system using integral images. In *In Proc. of the 5th International Conference on Computer Vision Systems (ICVS*, 2007.

[Frintrop *et al.*, 2010] Simone Frintrop, Ro Erich, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.*, 7:6:1–6:39, January 2010.

[Frintrop, 2006] Simone Frintrop. Vocus: A visual attention system for object detection and goal-directed search. *Lecture Notes in Artificial Intelligence (LNAI)*, 3899, 2006.

[Gao and Vasconcelos, 2004] Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In *In Proc. NIPS*, pages 481–488, 2004.

[Gao and Vasconcelos, 2005] Dashan Gao and Nuno Vasconcelos. Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 282–287, Washington, DC, USA, 2005. IEEE Computer Society.

[Gao *et al.*, 2008] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7), 2008.

[Geisler and Perry, 1998] Wilson S. Geisler and Jeffrey S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In *in Proc. SPIE: Human Vision and Electronic Imaging*, volume 3299, pages 294–305, 1998.

[Goferman *et al.*, 2010] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. In *CVPR'10*, pages 2376–2383, 2010.

[Grabli *et al.*, 2004] Stéphane Grabli, Frédo Durand, and François Sillion. Density measure for line-drawing simplification. In *Proceedings of Pacific Graphics*, 2004.

[Green and Swets, 1966] D. M Green and J. A Swets. *Signal detection theory and psychophysics.* John Wiley, 1966.

[Guo and Zhang, 2010] Chenlei Guo and Liming Zhang. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *Trans. Img. Proc.*, 19:185–198, January 2010.

[Harel *et al.*, 2007] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[Harmon and Julesz, 1973] Leon D. Harmon and Bela Julesz. Masking in visual recognition: Effects of two-dimensional filtered noise. *Science*, 180(4091):1194–1197, 06 1973.

[Henderson *et al.*, 1997] John Henderson, Karen Mcclure, Steven Pierce, and Gary Schrock. Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Attention, Perception, and Psychophysics*, 59:323–346, 1997. 10.3758/BF03211901.

[Henderson *et al.*, 1999] John M. Henderson, Phillip A. Weeks, and Andrew Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing, ,. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1):210 - 228, 1999.

[Henderson *et al.*, 2007] J M Henderson, J R Brockmole, M S Castelhano, and M Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye Movement Research: Insights into Mind and Brain*, 2007.

[Hoffman, 1998] J. E. Hoffman. Visual Attention and Eye Movements. In Harold Pashler, editor, *Attention*, pages 119–154. Psychology Press, 1998.

[Holtzman-Gazit *et al.*, 2010] Michal Holtzman-Gazit, Lihi Zelnik-Manor, and Irad Yavneh. Salient edges: A multi scale approach. In *ECCV International Workshop on Workshop on Vision for Cognitive Tasks*, 2010.

[Hou and Zhang, 2007] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[Itti and Baldi, 2006] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems, Vol. 19 (NIPS*2005)*, pages 547–554, Cambridge, MA, 2006. MIT Press.

[Itti and Koch, 2000] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[Itti and Koch, 2001] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2:194–203, 2001.

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

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

[Itti, 2005] Laurent Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12:1093–1123, 2005.

[James, 1890] William James. *The principles of psychology*, volume 2. Holt, 1890.

[Jost *et al.*, 2005] Timothe Jost, Nabil Ouerhani, Roman von Wartburg, and Rene M˙ Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107 – 123, 2005. Special Issue on Attention and Performance in Computer Vision.

[Judd *et al.*, 2009] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[Judd *et al.*, 2011] Tilke Judd, Fredo Durand, and Antonio Torralba. Fixations on low-resolution images. *Journal of Vision*, 11(4), 2011.

[Just and Carpenter, 1980] M.A. Just and P.A. Carpenter. A theory of reading: from eye fixation to comprehension. *Psychology Review*, 87:329–354, 1980.

[Kanan *et al.*, 2009] Christopher Kanan, Mathew H. Tong, Lingyun Zhang, and Garrison W. Cottrell. Sun: Top-down saliency using natural statistics, 2009.

[Kayser *et al.*, 2006] C. Kayser, K. J. Nielsen, and N. K. Logothetis. Fixations in natural scenes: interaction of image structure and image content. *Vision Res*, 46(16):2535–2545, August 2006.

[Kienzle *et al.*, 2006] Wolf Kienzle, Felix A. Wichmann, Bernhard Schölkopf, and Matthias O. Franz. A nonparametric approach to bottom-up visual saliency. In Bernhard Schölkopf, John C. Platt, and Thomas Hoffman, editors, *NIPS*, pages 689–696. MIT Press, 2006.

[Koch and Ullman, 1985] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurbiology*, 4:219–227, 1985.

[Kootstra *et al.*, 2008] Gert Kootstra, Arco Nederveen, and Bart De Boer. Paying attention to symmetry. In *Proceedings of the British Machine Vision Conference (BMVC2008*, pages 1115–1125, 2008.

[Le Meur *et al.*, 2006] Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:802–817, May 2006.

[Linde *et al.*, 2008] Ian Van Der Linde, Umesh Rajashekar, Alanc. Bovik, and Lawrence K. Cormack. Doves: a database of visual eye movements, 2008.

[Liu *et al.*, 2007] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, june 2007.

[Loftus and Mackworth, 1978] Geoffrey R. Loftus and Norman H. Mackworth. Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 4(4):565 – 572, 1978.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[Mannan *et al.*, 1997] S K Mannan, K H Ruddock, and D S Wooding. Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26(8):1059–1072, 1997.

[Mannan, 1995] S. Mannan. Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial Vision*, 9:363–386(24), 1995.

[Marchesotti *et al.*, 2009] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2232 –2239, 29 2009-oct. 2 2009.

[McCotter *et al.*, 2005] Maxine McCotter, Frederic Gosselin, Paul Sowden, and Philippe Schyns. The use of visual information in natural scenes. *Visual Cognition*, 12(6):938–953, 2005.

[Mertsching *et al.*, 1999] B. Mertsching, M. Bollmann, R. Hoischen, and S. Schmalz. The neural active vision system navis. In *In Handbook of Computer Vision and Applications, B. Jahne, H. Haussecke, and P. Geissler, Eds.*, volume 3, page 54356, 1999.

[Meur *et al.*, 2007a] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[Meur *et al.*, 2007b] Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[Miau *et al.*, 2001] F. Miau, C. Papageorgiou, and L. Itti. Neuromorphic algorithms for computer vision and attention. In B. Bosacchi, D. B. Fogel, and J. C. Bezdek, editors, *Proc. SPIE 46 Annual International Symposium on Optical Science and Technology*, volume 4479, pages 12–23, Bellingham, WA, Nov 2001. SPIE Press.

[Milanese, 1993] R Milanese. *Detecting salient regions in an image: from biological evidence to computer implementation.* PhD thesis, 1993.

[Navalpakkam and Itti, 2005] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205 – 231, 2005.

[Navalpakkam and Itti, 2006] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2049–2056, New York, NY, Jun 2006.

[Navalpakkam and Itti, 2007] Vidhya Navalpakkam and Laurent Itti. Search goal tunes visual features optimally. *Neuron*, 53:605–617, 2007.

[Neider and Zelinsky, 2006] Mark B. Neider and Gregory J. Zelinsky. Scene context guides eye movements during visual search. *Vision Research*, 46(5):614 – 621, 2006.

[Oliva and Schyns, 2000] Aude Oliva and Philippe G. Schyns. Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41(2):176 – 210, 2000.

[Oliva and Torralba, 2001a] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, May 2001.

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

[Oliva *et al.*, 2003] A. Oliva, A. Torralba, M.S. Castelhano, and J.M. Henderson. Top-down control of visual attention in object detection. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, volume 1, pages I – 253–6 vol.1, sept 2003.

[Oliva, 2005] Aude Oliva. Gist of the scene. *Neurobiology of Attention*, page 251256, 2005.

[Ouerhani, 2003] Nabil Ouerhani. Visual attention: From bio-inspired modeling to real-time implementation. ph.d. thesis, institut de microtechnique universit de neuchtel, switzerland, 2003.

[Parkhurst and Niebur, 2003] D.J. Parkhurst and E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16:125–154(30), 2003.

[Parkhurst *et al.*, 2002] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107 – 123, 2002.

[Peters *et al.*, 2005] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397 – 2416, 2005.

[Pinto *et al.*, 2009] Nicolas Pinto, David Doukhan, James J. DiCarlo, and David D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput Biol*, 5(11), 11 2009.

[P.J. Lang and Cuthbert, 2008] M.M. Bradley P.J. Lang and B.N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual., 2008.

[Potter and Levy, 1969] Mary C. Potter and Ellen I. Levy. Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, 81(1):10–15, 1969.

[Potter, 1975] MC Potter. Meaning in visual search. *Science*, 187(4180):965–966, 1975.

[Privitera and Stark, 2000] Claudio M. Privitera and Lawrence W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:970–982, September 2000.

[Rajashekar *et al.*, 2008] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack. Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, 17(4):564 –573, april 2008.

[Ramanathan *et al.*, 2010] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *ECCV 2010*, Crete, Greece, 2010.

[Renninger *et al.*, 2007] Laura Walker Renninger, Preeti Verghese, and James Coughlan. Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 2007.

[Riesenhuber and Poggio, 1999] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

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

[Rothenstein and Tsotsos, 2008] Albert L. Rothenstein and John K. Tsotsos. Attention links sensing to recognition. *Image and Vision Computing*, 26(1):114 – 126, 2008. Cognitive Vision-Special Issue.

[Rubinstein *et al.*, 2008] Michael Rubinstein, Ariel Shamir, and Shai Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics (SIGGRAPH)*, 2008.

[Rubner *et al.*, 2000] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.

[Russell *et al.*, 2005] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. Labelme: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, MIT CSAIL, September 2005.

[S. *et al.*, 1992] Hacisalihzade S. S., Allen J.S., and Stark L. Visual perception and sequences of eye movement fixations: A stochastic modelling approach. *IEEE Transactions on Systems, Man and Cybernetics*, 22:474–481, 1992.

[Santella *et al.*, 2006] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, New York, NY, USA, 2006. ACM.

[Schyns and Oliva, 1997] P G Schyns and A Oliva. Flexible, diagnosticity-driven, rather than fixed, perceptually determined scale selection in scene and face recognition. *Perception*, 26(8):1027–1038, 1997.

[Seo and Milanfar, 2009a] Hae Jong Seo and P. Milanfar. Nonparametric bottom-up saliency detection by self-resemblance. *Computer Vision and Pattern Recognition Workshop*, 0:45–52, 2009.

[Seo and Milanfar, 2009b] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 2009.

[Shipp, 2004] Stewart Shipp. The brain circuitry of attention. *Trends in Cognitive Sciences*, 8(5):223 – 230, 2004.

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

[Simoncelli and Freeman, 1995] E.P. Simoncelli and W.T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *ICIP95*, pages III: 444–447, 1995.

[Simoncelli, 2011] Eero Simoncelli. The steerable pyramid toolbox. http://www.cns.nyu.edu/ eero/steerpyr/, 2011.

[Sinha *et al.*, 2006] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948 –1962, nov. 2006.

[Sonnenburg *et al.*, 2006] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

[Suh *et al.*, 2003] Bongwon Suh, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs. Automatic thumbnail cropping and its effectiveness. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, UIST '03, pages 95–104, New York, NY, USA, 2003. ACM.

[Tatler and Vincent, 2009] Benjamin W. Tatler and B.T. Vincent. The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17(6-7):1029–1054, 2009.

[Tatler *et al.*, 2005] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643 – 659, 2005.

[Tatler, 2007] Benjamin W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14), 2007.

[Torralba and Efros, 2011] Antonio Torralba and Alexei Efros. Unbiased look at dataset bias. In *In proceedings of CVPR*, 2011.

[Torralba *et al.*, 2006] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, October 2006.

[Torralba, 2009] Antonio Torralba. How many pixels make an image?, 2009.

[Treisman and Gelade, 1980] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[Tseng *et al.*, 2009] Po-He Tseng, Ran Carmi, Ian G. M. Cameron, Douglas P. Munoz, and Laurent Itti. Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, 9(7), 2009.

[Tsotsos *et al.*, 2005] John K. Tsotsos, Laurent Itti, and G Rees. A brief and selective history of attention. *Neurobiology of Attention*, 2005.

[Valenti *et al.*, 2009] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *IEEE International Conference on Computer Vision*, 2009.

[Velichkovsky *et al.*, 1996] B. M. Velichkovsky, Marc Pomplun, Johannes Rieser, and Helge J. Ritter. *Attention and Communication: Eye-Movement-Based Research Paradigms.* Visual Attention and Cognition. Elsevier Science B.V., Amsterdam, 1996.

[Vig *et al.*, 2010] Eleonora Vig, Michael Dorr, Thomas Martinetz, and Erhardt Barth. A learned saliency predictor for dynamic natural scenes. In *Proceedings of the 20th international conference on Artificial neural networks: Part III*, ICANN'10, pages 52–61, Berlin, Heidelberg, 2010. Springer-Verlag.

[Viola and Jones, 2001] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[Walther and Koch, 2006] Dirk Walther and Christof Koch. Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395 – 1407, 2006. Brain and Attention, Brain and Attention.

[Wang and Li, 2008] Zheshen Wang and Baoxin Li. A two-stage approach to saliency detection in images. *IEEE International Conference on Acoustics Speech and Signal Processing ICASSP*, pages 965–968, 2008.

[Wolfe *et al.*, 1989] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, Vol 15(3):419–433, 1989.

[Wolfe, 1998] Jeremy M Wolfe. Visual memory: What do you know about what you saw? *Current Biology*, 8(9):R303 – R304, 1998.

[Yarbus, 1967] A. L. Yarbus. Eye movements and vision. 1967.

[Zhang *et al.*, 2008] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008.

[Zhao and Koch, 2011] Qi Zhao and Christof Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.