Abstract

Computational Methods for Eye-Tracking Analysis: Applications to Autism

Frederick Shic

2008

Though eye-tracking technology has developed considerably over the last hundred years, eye-tracking analysis is still in its infancy. This thesis describes computational techniques and methods we have developed for augmenting this analysis. These methods correct deficits in current approaches, extend traditional techniques to gain greater clarity, and provide frameworks for viewing gaze patterns from new perspectives. We use our methods to study autism, a disorder characterized by social and communicative deficits, in order to gain insight into how these individuals attend to the world around them, and to determine what factors may be motivating their attention.

We begin by showing how current fixation algorithms for eye-tracking analysis provide an incomplete picture of gaze behavior. We present a simple linear interpolation model (SLIM) that can provide a more complete, but still compact, picture. We apply this model to the scanning patterns of toddlers with autism and show results which coincide with known deficits in face processing. Furthermore, by adapting standard fixation algorithms to perform temporally greedy box-counting, we provide evidence that

the incompleteness of standard algorithms may be due to the fractal qualities of the underlying scanning distributions.

Examining distributional aspects of scanning provides only an overview of differences. For this reason we examine standard, fine-grained, region-of-interest (ROI) eye-tracking analysis where regions are drawn around areas and measures, such as how long a subject looks at areas, are calculated. Typically, dynamics of scanning are ignored. To correct this, an entropy measure, as an index of exploration, is proposed and applied to children with autism. We show a pervasive pattern of inattention in autism differentiating 4 year old, but not 2 year old, children with autism from typical children, and discuss how atypical experience and intrinsic biases may affect development.

As an alternative to ROI analysis, which can be a subjective and laborious top-down approach, a bottom-up evaluation, based on computational modeling of low-level features, is offered. We use these models to examine preferences for low-level features in autism, and show that children with autism attend more to areas of contrast and less to areas of motion. We also use these same models for gauging the gaze distance between individuals. We use these techniques to highlight the heterogeneity of autism, showing how gaze patterns of individuals with autism are as different from each other as they are from typical controls, and discuss the factors which might lead to this heterogeneity.

Finally, we conclude with a discussion of the advantages of the methodologies that we have presented, and discuss the results of our work as they pertain to both the computational and methodological advances we have accomplished and the insights that we have obtained regarding autism.

Computational Methods for Eye-Tracking Analysis: Applications to Autism

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Frederick Shic

Dissertation Director: Brian Scassellati
December 2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

Contents

Figures

Tables

Acknowledgments

I would like to thank my advisor, Brian Scassellati, for showing me how to solve difficult problems simply, and for helping me to understand how to find those problems worth solving. He has given me support when I have needed, has helped me find my way back when I was lost. Most importantly, he cured me of a disease whereby I would use a genetic algorithm to solve everything. I would also like to thank my mentor in psychopathology, Katarzyna Chawarska, who has truly been a mentor in every sense of the word. It is a testimony to her patience, and willingness to nurture, that a computer scientist can sit at a table with clinical psychologists and actually understand what they are saying, as long as they're not talking about Freud.

I would also like to thank Steven S. Zucker and John Tsotsos for agreeing to read my thesis. Having them both on my committee is no accident. I hold them as models of both scientists and individuals; both have affected my thinking and my work, I expect, more than they know. It is with an apology that I hand them this thesis, which is 100 pages longer than they expected. I would also like to thank the many excellent faculty who have taught me and guided me, especially Drew McDermott and Willard Miranker, for allowing me to teach their classes. Those students never knew what hit them. And I would especially like to thank Arvind Krinshnamurthy who has been as much a friend as he has been a teacher.

I would especially like to thank Ami Klin, Warren Jones, and Fred Volkmar, for first inspiring in me an interest in autism research. It is likely that, had we never met, this

thesis would have been about genetic algorithms. I would also like to thank the faculty and staff I have met over these years at the Yale Child Study Center: Joe Chang, David Lin, Suzanne Macari, Linda Mayes, Jamie McPartland, Rhea Paul, and Gordon Ramsey. Also, this thesis would not have been possible if not for the brilliant work of the research assistants at the Child Study Center: Jessica Bradshaw, Brittany Butler, Rebecca Doggett, Sarah Hannigen, Joslin Latz, Allison Lee, Paula Ogston, and Jessica Reed.

This work was supported by NIH Research Grant U54 MH66494 funded by the NIMH, NIDCD, NIEHS, NICHD, and NINDS; NSF CAREER award (#0238334) and NSF award #0534610; a software grant from QNX Software Systems Ltd.; support from the Sloan Foundation; and an Autism Speaks mentor-based pre-doctoral fellowship. Thank you all for making me feel like a pro athlete, except not as well paid, and without any physical skill.

I would like to thank my lab mates at the Yale Social Robotics Lab who have been both my friends and my colleagues: Christopher Crick, Marek Doniec, Kevin Gold, Justin Hart, Eli Kim, Marek Michalowski, Philipp Michel, and Ganghua Sun. I would like to thank my office mates, who had to deal with my perpetually obnoxious schedule and constant messiness: Hao Wang and Yinghua Wu. And, of course, to my friends who have made these years as a graduate student bearable and often fun (special props for John Corwin for his instruction in lethality; Kevin Chang for the constancy of his companionship (and numerous diversions)): thank you!

To my wife, Annie, who has endured all my foibles and weaknesses, has stood by me in all my toils, and is my sunlight, even when the day starts at 9PM: you are everything I could ever want, and more than I could ever deserve. To my daughters,

Adenine and Tesla, you are the joy in my life, and I am so, so very sorry for giving you those names. You are as special to me as the sun is to the earth, only brighter, because you're mine.

To my mother, who always tackles every problem with playfulness and delight, ever curious, and who is more imaginative and more capable than anyone knew: you are an inspiration to me. You always wondered what I've been doing: here it is. Finally, to my father, my model for strength and perseverance, who passed away in my first semester of my graduate school: I love you and miss you. We got it done.

It is to the parents who gave me life, and a life worth living, that I dedicate this thesis.

Chapter 1

Introduction

It has been said that "the eyes are a window into the soul". Since the inception of eye-tracking more than a century ago (Huey, 1898), researchers have been seeking to make this proverb explicit by mapping the movements of the eyes to the motives of the individual. In this work, we develop computational and analytic techniques for accomplishing this mapping, examining individuals with autism spectrum disorders (ASD) in order to refine our methodological approaches as well as to elucidate the condition itself.

Autism is a pervasive developmental disorder marked by severe deficits in social functioning (American Psychiatric Society, 1994). We choose to examine individuals affected with autism and individuals with related conditions (i.e. the "spectrum" of autistic disorders) for two main reasons. First, we know from a host of studies and reports that individuals with ASD do not necessarily view the world in the same way that typical individuals do (e.g. see Baron-Cohen, 1995b; Frith, 2003a; Grandin, 1992; Happé, 1999b; Lawson, 2001; Mayes & Cohen, 1994; Ozonoff, Pennington, & Rogers, 1991). Thus, it would seem likely that the study of autism with eye-tracking, especially in less constrained, more natural experimental paradigms such as free-viewing (i.e. viewing of scenes without explicit instructions), would be able to reveal striking differences between typical individuals and individuals with autism. The second reason is practical: autism is one of the most common developmental disorders affecting children. Recent estimates suggest that about 1 in 150 children are affected with ASD, making autism more common

than Down syndrome, juvenile diabetes, and childhood cancer (Center for Disease Control and Prevention, 2008). Given the potential for early treatment in autism (Bryson, Rogers, & Fombonne, 2003; Goldstein, 2002; Green, Brennan, & Fein, 2002; Lovaas, 1987; McEachin, Smith, & Lovaas, 2001; Rogers, 1998; Sheinkopf & Siegel, 1998; Smith, 1999), work that brings us closer to better quantitative measures for early diagnosis, or provides for us a greater understanding of the factors affecting the developmental progression of autism, is especially critical.

It is important to note that though many of our experiments will examine individuals with autism, the methodologies that we will present are not limited to the study of autism, but have many applications to a wide variety of fields. Every effort will be made so that discussions regarding the methodologies themselves, which is truly the focus of our work, is as distinct as possible from the discussions regarding autism. However, it is also the case that it is often difficult to disentangle method from application, and we hope that, through example, we can convey the need for tailoring models, techniques, and general approaches to the needs of the specific investigation.

We begin by asking, "how can we understand how a person views the world?" For a typical individual, the simplest approach would be to ask him. However, it is well known that introspection can be a problematic method for uncovering internal mental states (Frith & Lau, 2006; Nisbett & Wilson, 1977; Overgaard, 2006). In fact, though the eye jumps around quickly in rapid movements called saccades, individuals are typically not even aware that their eyes are moving due to active suppression of this awareness by the brain (Burr, Morrone, and Ross, 1994; Erdmann and Dodge, 1898); instead the visual world is perceived as a coherent whole. In addition, asking someone how he perceives

the world can only be effective if the subject can understand the questions asked of him. For this reason, this method of asking someone what he sees is controversial and difficult to use in young children (Estes, 1994) and with individuals from atypical populations, such as autism, who are either impaired in their ability to comprehend language or social norms (Frith & Happé, 1999; Baron-Cohen, 1995b).

For these reasons eye-tracking has become one of the most popular methods for uncovering what it is that individuals actually see (for a survey of applications, see Duchowski, 2002). There is ample evidence that suggests that a strong link exists between cognition and the focal point of gaze. Some of the first systematic studies of scene perception showed that individuals viewing artwork did not sample from a scene uniformly, but rather skipped over uninformative regions in favor of content-rich areas (Buswell, 1935). Later, Yarbus (1967) showed that the eye scanning patterns of individuals viewing pictures changed when they were given different sets of instructions, helping to establish the dependence of eye movements on the internal goals and motivations of the individual. More recent work has further elucidated these relationships and confirmed the dependence of eye-movements on an individual's moment-by-moment needs (for reviews, see Hayhoe & Ballard, 2005; Henderson & Hollingworth, 1999; Henderson, 2003). For example, when engaging in everyday tasks, such as making a sandwich, a person will focus on objects and tools as they are needed (Ballard, Hayhoe, & Pelz, 1995; Hayhoe, 2000; Hayhoe & Ballard, 2005; Land & Hayhoe, 2001).

The strong link between gaze and cognition has lead to the development of several theories for describing the mechanisms behind this link. However, these theories

have not been without controversy. Noton and Stark (1971) proposed what has become to be known as the "scanpath theory", the suggestion that the physical pattern the eye takes when viewing a scene is integrally tied to the encoding and retrieval of that scene in memory. This theory was later shown to be somewhat overoptimistic, as individuals can identify scenes previously viewed with just one fixation, and because the pattern of fixations which make up the scanpath is actually highly variable (for a short discussion see Henderson, 2003). Just and Carpenter (1980) proposed a theory based on two assumptions: 1) "immediacy", that information begins to be processed the moment it is fixated upon, and; 2) the "eye-mind assumption", that the eye remains fixed on a target as long as it is being processed. However, it is known that it is possible to keep the eyes still while covertly shifting attention (see Posner, 1980). This process of covert attention would suggest that information can be processed before it is fixated, and also suggests that the duration of a fixed eye gaze does not necessarily need to be exactly matched to the processing time (Inhoff, Pollatsek, Posner, & Rayner, 1989).

However, though some studies might suggest that eye-tracking reflects where an individual's eyes go, and not necessarily what they perceive, in more natural, less artificial, experimental settings, other studies have shown that when the eyes move, attention tends to follow (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Kowler, Anderson, Dosher, & Blaser, 1995; Shepherd, Findlay, & Hockey, 1986). Thus, while it may not be reasonable to assume that the eye and the cognition are perfectly matched, in most situations the assumption that the two are highly correlated is a fair one.

In autism, there are many reasons to believe that the use of eye-tracking might be especially efficacious (Klin, Jones, Schultz, Volkmar, & Cohen, 2002b). For instance, it

has been hypothesized that the social dysfunction evident in autism is the result of an early derailment of the typical experience-dependent social-cognitive developmental process (Klin, Jones, Schultz, Volkmar, & Cohen, 2002a). Evidence for this theory is provided for by the atypical looking patterns of children with autism viewing naturalistic dynamic scenes (Figure 1.1). In contrast to typical controls, individuals with autism attend preferentially to mouths and bodies of characters rather than eyes (Klin et al., 2002a). As the eyes of an individual convey a great deal of information about his or her internal mental state (Baron-Cohen, Campbell, Karmiloff-Smith, Grant, & Walker, 1995), not looking at the eyes would necessarily lead to deficits in processing social information.

It has also been hypothesized that the established atypical viewing patterns have some neural basis, which is to say that abnormal looking patterns are not the ultimate cause of social dysfunction, but rather an expression of some underlying neurocognitive divergence (Belmonte et al., 2004). Insight into a neurocognitive mechanism is potentially uncovered by exploring basic perceptual abnormalities in individuals with autism. These perceptual abnormalities could bias the child with autism away from building the typical scaffolding upon which social skills are built. For example, individuals with autism are known to have preferences and advantages for local visual processing (as compared to global processing) (Frith, 2003b; Happé, 1999b; Rinehart, Bradshaw, Moss, Brereton, & Tonge, 2000). This inherent preference may play a role in discrepancies observed during the viewing of inverted faces: whereas typical individuals are disturbed by inversion (likely due to disruption of global configural features), individuals with autism are not (Tantam, Monaghan, Nicholson, & Stirling, 1989). In

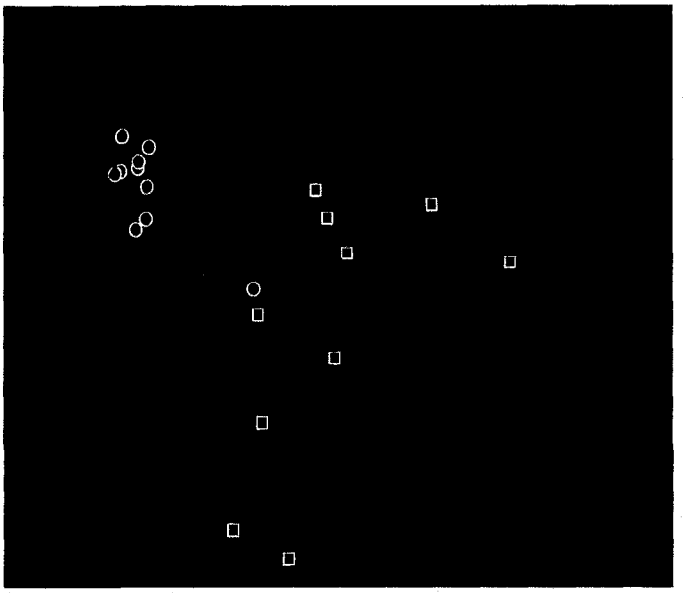


Figure 1.1: Eye scanning paths of controls (solid lines from circles) as compared to individuals with autism (dotted lines from squares) on a scene from the 1966 movie "Who's Afraid of Virginia Woolf?" (Klin et al., 2002a). The instantaneous fixation point is the circle or square and each path stretches 250 ms into the future. The gaze locations of controls are clustered on the left-most face; the gaze locations of individuals with autism are scattered.

addition, there is evidence for motion processing deficits in autism (Dakin & Frith, 2005; Milne et al., 2002) as well as the lack of salient attribution to biological motion (Blake, Turner, Smoski, Pozdol, & Stone, 2003).

These last two paragraphs give us a notion of why eye-tracking in autism has such great potential. Autism is a complex and multifaceted disorder. The social dysfunction is central, so how do these other factors, such as local processing, invariance to face inversion, motion processing deficits, relate? The etiology of autism is not yet well understood, nor all the developmental trends. To investigate these aspects, one should consider not just adults with autism, but also children. And if social deficits are the

central dysfunction, it becomes more difficult to rely on the ability to successfully communicate instructions for a test. On the other hand, eye-tracking is a methodology that can be applied in non-verbal or mentally disadvantaged individuals, from infants to adults. It is a remarkably powerful tool, truly a technology that can give us a window into the minds of these children -- provided that we are looking in the right place.

However, it is one thing to say that the movement of the eyes depends on the cognitive state of the individual, but quite another to say the converse, i.e. that it is possible to decode the cognitive state of an individual from the movement of his eyes. Eye-tracking, especially when studying how individuals view scenes without explicit instruction, suffers from the same difficulties as traditional psychological studies of looking time. Notably, there exists what Aslin (2007) calls the "many-to-one mapping problem": many different factors could be responsible for observed effects, especially as the complexity of the presented stimuli, and thereby the number of possible compounding factors, increases. It is thereby necessary to 1) constrain experiments so that the effects of confounding factors is a minimum; and 2) to develop multiple methods for modeling the observed effects, so that the observed phenomena can be viewed from multiple perspectives. The first point is, of course, a general principle, and should be followed in any good experimental paradigm. We have addressed this point mainly by focusing on between subject variation, i.e. the comparison of individuals with autism against typical individuals. The second point is the primary focus of this work. We present a variety of novel techniques and approaches which we can use to better view the gaze patterns of individuals from multiple angles, overcoming many of the limitations and deficits of the current practices in eye-tracking analysis.

We begin with a brief review of the different types of eye-tracking technology in Chapter 2. In all of our work, we use one particular type of eye-tracker, and thus it is useful to understand both the mechanics and limitations of the hardware and the processing that provide a front-end for all our research. We will then discuss the eye-tracking processing pipeline beginning from the point where the eye-tracker detects markers on the eye that are to be tracked. In turn we will discuss blink detection, data quality measures, calibration of the eye-tracking system, and fixation identification. We will devote a large amount of time to fixation identification, which is the process of separating fast, ballistic motions of the eye (saccades) from the periods of time where the eye is relatively stable upon the scene (fixations).

In Chapter 3, we will examine some of the assumptions that are made in eye-tracking, focusing specifically on fixation identification. We will examine one common measure that is used on fixations, the mean fixation duration, and show that the traditional interpretation of this measure is incomplete. We will refute the blind assumption that this measure has some physiological analogue. We will further develop models which, though quite simple, seem to provide a more complete picture of subject behavior during our experiments. At the end of this chapter we will present some preliminary evidence that what is tapped by standard fixation identification algorithms is a small part of a richer spatiotemporal distribution, and suggest there may be a fractal-like structure to gaze-patterns, raising further questions as to the appropriateness of standard fixation analysis.

In Chapter 4 we will discuss the traditional method for analyzing gaze patterns: region based analysis. In this form of analysis, the scene that is presented to a subject is

divided into regions and measures over eye-tracking data within these regions are interpreted. Typically, the dynamic component of scanning, such as transitions between regions, is ignored. We will thus provide an entropy measure as an index of exploration which can be used for tapping higher-order transitional effects in eye-tracking data. In addition, we will provide some guidelines as to how to manage the complexity of high-level eye-tracking analysis.

In Chapter 5 we will move into the realm of computational modeling and discuss the distinction between models of visual attention that are meant to provide predictions as to where eye-patterns should go (predictive models), and models of visual attention that are meant to provide descriptions or explanations as to the underlying structure of scanning (descriptive models). We will organize a framework for describing computational models of visual attention and discuss their components. We will give some examples of these computational models, especially regarding one popular biologically-inspired model of visual attention, that of Itti et al. (1998). As we are interested in using these computational models to examine scanning patterns on dynamic scenes, we augment the model of Itti et al (1998), which was originally intended to operate only on still images, with a motion extension which is more in line with the basic spirit of the Itti model than other work. We will return to this model several times in our subsequent work.

In Chapter 6 we will address the question, "how do we know if two gaze patterns are the same or different?" This is a difficult question because the scene coming in to the retina changes as the eye moves, and the effects are even more complex when considering time-varying scenes. We will answer this question by grounding the gaze

patterns in the features of the scenes under view. Furthermore, we will show how different computational models of visual attention can be evaluated against each other in a behavioral sense by comparison against human observers. We will develop a strategy which will tune computational models to gaze patterns, thus providing a more level ground for the comparison of these models. We will use this methodology to evaluate several different models of visual attention and show that bottom-up models of visual attention do not necessarily have an advantage in the predictive sense. We will illustrate this point with experiments on the model of Itti et al. (1998).

In Chapter 7 we will show how the general framework developed in Chapter 6 can be used to measure how well one person's model of gaze describes the gaze patterns of others. This will lead to a natural method for between-subject comparisons. When combined, we show that predictive models can offer some insight into the dynamics of different subject populations.

Finally, in Chapter 8, we will move away from predictive models and back towards descriptive models. We will take the model of Itti et al. (1998) and strip away the predictive aspects, and, by using the core components of the model, extract a bottom-up interpretation for gaze analysis. We use this bottom-up interpretation in order to examine different subject populations under different types of contextual modulation, and show how computational models of visual attention can provide advantages for interpretation and evaluation, even when they do not necessarily show an advantage in the predictive sense.

As we have mentioned, along the way, we will apply our techniques and methods to the gaze patterns of either children or adults with autism spectrum disorders. Though

this thesis deals primarily with computational methodology, just as eye-tracking on its own can tell us where someone is looking but not why, the methodology is only a description until we tackle hard problems with it.

Chapter 2

Processing Eye-Tracking Data

The initial stages of processing eye-tracking data are topics that are not usually broached in the discussion of new approaches for eye-tracking analysis. Typically, these initial stages are handled invisibly by the manufacturers of the eye-tracking systems so that the end user only has to work in the "world coordinates" of the stimuli presented to subjects (e.g. in the pixel coordinates of a computer display). However, the end results of any analysis can be heavily influenced by assumptions inherent in the beginning stages of the analytical pipeline. In this chapter we will discuss how eye-tracking data is typically processed before analysis is initiated. We will see that many of the techniques used in common practice are simple heuristics, and in the next chapter we will see how one deeply held assumption, that of the appropriateness of fixation identification, is flawed, and how this single simple assumption renders the use of a traditional eye-tracking measure questionable. It is important to note that the full analysis of all the heuristics used in the initial stages of eye-tracking processing is beyond the scope of this work, but the reader is encouraged to consider the ramifications of each of the assumptions used in the front end of eye-tracking processing as they are encountered. It is likely that future work will make improvements on the heuristics discussed, increasing the accuracy of the resultant scanning patterns and providing a better foundation for analysis and interpretation.

To understand the current limits and tradeoffs in eye-tracking technology, we will begin with a discussion regarding some of the various types of eye-tracking systems that

are widely used in psychological and cognitive science research today. We will continue with a discussion regarding the processing of data from one of the most adaptable techniques in use today, table-mounted video-based pupil/corneal reflection systems, and describe the processing pipeline beginning from after pupil and corneal reflection locations are localized. In turn, we will discuss eye-tracking calibration, accuracy measurements, blink detection, and fixation and saccade identification.

2.1 Types of Eye-tracking

Eye-tracking is a very old technology (Huey, 1898) and today there are several types of eye-tracking systems in common use (for a brief survey of eye-tracking systems see Duchowski (2003)). Electro-oculography (EOG) is a technique by which electrodes are placed on the skin around the eyes and the difference in surface potential (accessing the resting potential of the retina) is used to calculate the current position of the eye. The advantages of EOG include its cheap cost, but the disadvantages include a relatively poor spatial (~1-2°) and temporal (~40hZ) resolution (Hain, 2008), drift due to changes in skin conductance, and the fact that the head position must be either tracked or constrained in order to obtain reliable results (Heide & Zeec, 1999). Furthermore, the mounting of electrodes on the skin, though well tolerated by typical adults, is somewhat time-consuming and not always tolerated by certain subject populations, such as children or individuals with mental disorders.

Another common technique, still widely used in animal research, is the scleral coil technique. In this technique a contact lens is mounted on the eye together with a

reference object, such as a wire. As the eye moves, so does the scleral coil. Scleral coils are amongst the most accurate and precise eye-tracking devices, with excellent spatial (~.03°) and temporal resolution (1 kHz typical) (Roberts, Shelhamer, & Wong, 2008). Unfortunately, if EOG electrodes are poorly tolerated by some individuals, scleral coils are poorly tolerated by all individuals. The coil must be delicately positioned and often the eye must be anesthetized; there is also a risk of corneal abrasion. Very recently it was determined that the wire leading off wire-based scleral systems was the leading cause of discomfort in subjects wearing them and a radio-wave resonating wireless coil system was developed and found to greatly reduce irritation (Roberts et al., 2008). However, it is likely that these systems will have limited applicability to special populations, such as infants, as it still involves the insertion of a contact lens into the eye. As with EOG, care must be taken with scleral coils in order to guarantee that the head is immobilized or tracked in order to determine the subject's point of regard.

The most common eye-tracking systems in use today on human subjects are video-based pupil/corneal reflection eye-tracking systems. These systems rely on video localization of the pupil in conjunction with infrared illumination. The infrared illumination reflects off the cornea and the location of these corneal reflections can then be detected and used as a benchmark to gauge the relative position of the pupil, making this type of eye-tracking system resilient to subject motion. Spatial resolution is better than most EOG systems (.1-1°) and adequate for most applications. Temporal resolution depends on the camera frame rate: 50Hz and 60Hz systems are common, with higher speed systems (250Hz+) finding more applications as high-speed digital camera costs decrease. There are many variants of these video-based systems, including both head-

mounted and table-mounted eye-trackers. Head-mounted systems suffer from some of the same invasive and obtrusive characteristics as EOG and scleral systems, being physically mounted on the subject's head, but typically take less time for preparation and are less intrusive. Table-mounted systems are the least invasive, though some care still must be taken for identifying and compensating for subject motion. Because an individual can sit in a chair with no attachments to his body whatsoever, table-mounted video-based systems are the preferred choice in experimental protocols where subjects are sensitive to touch or might otherwise mishandle or damage eye-tracking equipment in close proximity. In addition, table-mounted video eye-tracking offers the most natural environment in which to perform experiments, as there are no constant physical reminders, other than a relative immobility (i.e. the fact that the subject must sit in a chair), that an experiment is taking place.

Thirty years ago, eye-tracking systems were confined to select research institutions. This is no longer the case. The advent of cheap, commercialized eye-tracking systems has lead to a proliferation of eye-tracking research in a wide range of domains, from psychology and neuroscience to computer interaction and marketing research. However, whereas before scientists built their own eye-tracking systems and were all experts in eye-tracking, today, in many cases, the end-users of eye-tracking systems are unaware of the internals of the eye-tracking systems and the extensive processing necessary to bring eye-movement data into a useful form for analysis. It is hoped that this section helps to remedy this situation by presenting the details of the front-end processing of eye-tracking systems. Furthermore, the latter part of this chapter

will provide a background to the subsequent investigations of the validity of the assumptions by which eye-tracking analysis is based.

As many of the algorithmic specifics are kept proprietary by eye-tracker manufacturers, we have had to reengineer much of the eye-tracking pipeline. Research by Duchowski (2003) and Salvucci & Goldberg (2000) were useful references in this process. In this chapter we will restrict our attention to one of the most popular types of eye-tracking environments found in psychological and cognitive research: table-mounted video-based pupil/corneal reflection eye-trackers operating at normal speeds (60Hz) for tracking on a 2D environment. With some modification, many of the techniques we will describe in the following sections would also be relevant to other eye-tracking systems. In addition, the perspective of much of this work is geared towards eye-tracking as a post-hoc analytical tool and, as such, the requirements of real-time performance are not our primary focus. Though many of our algorithms are quite efficient, some aspects, such as our use of bootstrapping for calibration and our later full-coverage fixation identification analysis, could be prohibitively expensive if real-time constraints were demanded. Still, with computing power increasing daily, what cannot run in real-time today might tomorrow. It is hoped that aspects of these techniques will find broader application as hardware and algorithms improve over time.

2.2 Features for Video-Based Eye-Tracking

A typical setup (SensoMotoric Instruments (SMI), 2006) consists of a camera focused on the area around the eye of the subject and one or more infrared sources for creating

corneal reflections (Figure 2.1). Note that the infrared sources shown in Figure 2.1 are actually black (dimly red in low-light conditions), and the purple glow is not visible to the human eye, but the digital camera used to record the scene has a sensitivity range that extends further into the infrared spectrum.

Typically, the eye camera system is connected to a computer which localizes the pupil and corneal reflections in the incoming video stream. For wide-view video systems (e.g. systems which record the head and/or torso of a subject), there are several methods for finding faces in images (Hjelmas & Low, 2001; Rowley, Baluja, & Kanade, 1998; Schneiderman & Kanade, 2000; Viola & Jones, 2004), and this typically aids in subsequent localization of the eyes by restricting the search domain for subsequent eye detection (Chow & Li, 1993; Cristinacce & Cootes, 2003; Jeng, Liao, Han, Chern, & Liu, 1998; Saber & Murat Tekalp, 1998) . Note that since it is typically a valid assumption that the face of the subject is frontal-facing and centered in the incoming video stream, many simplifications can be applied and general template-based or rule-based algorithms (Brunelli & Poggio, 1993; Duda, Hart, & Stork, 2000; Lewis, 1995) should work well. Similarly, in many systems only the eye is localized, further improving resolution and simplifying template or rule-based segregation of pupil and corneal reflections. In Figure 2.2 we illustrate simple rule-based detection of the pupil center and the corneal reflections on a eye-only infrared video stream. The pupil was detected by locating dark regions in the eye. This was followed by region growing whereupon the largest region was selected as the pupil. The centroid of this mass was taken as the pupil center. The corneal reflections were detected by convolving the image with a difference of Gaussians filter (σ_{center} = 1 pixel; $\sigma_{surround}$ = 2 pixels; filter side length = 4σ) (Lowe, 1999) and

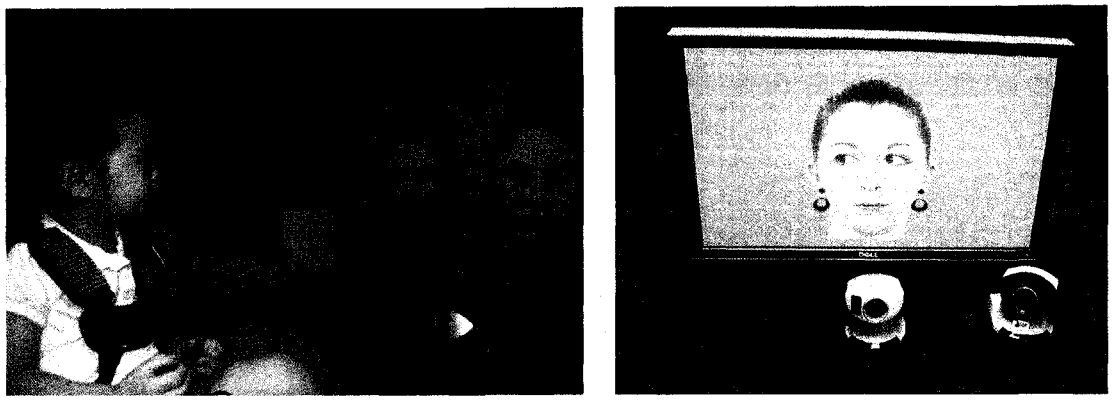


Figure 2.1: Video-based eye-tracking. *Left*: a child in an eye-tracking experiment (Singer-Vine, 2008). *Right*: the eye-tracking system. The infrared camera at the right side of the screen records the eye and the camera in the center records the entire face (for later coding and data-verification purposes). The two purple half-disks to the left and right of the monitor are infrared light sources.

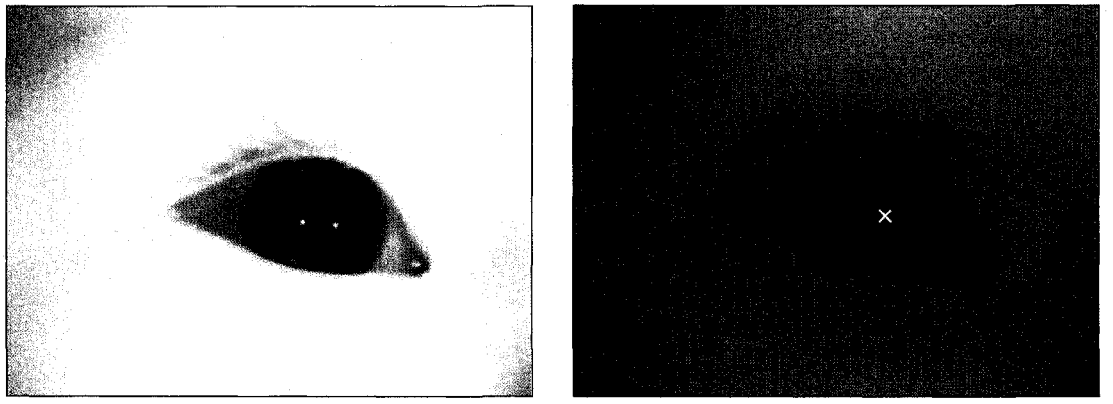


Figure 2.2: Rule-based Pupil/Corneal Reflection Detection. *Left*: image from the eye camera (Singer-Vine, 2008). *Right*: simple rule-based heuristics localizing the pupil (red), pupil-center (yellow x), and centers of corneal reflections (cyan cross).

looking for the brightest locations. Again region growing was used to identify corneal reflection regions and the centroid was used as the representative location of the reflections.

The input stream of data from the eye-tracker typically consists of four parts which vary with time (t):

1) the location of the center of the pupil, $\boldsymbol{r}'(t)$
2) the location of one or more corneal reflections $\boldsymbol{c}'_i(t)$, $i \in \{1..N\}$, for N corneal reflections
3) the vertical and horizontal size of the pupil $\boldsymbol{d}'(t) = (d'_{horizontal}(t), d'_{vertical}(t))$
4) an initial Boolean flag for each data stream indicating whether data on that stream was valid or invalid at that time (i.e. whether the eye-tracker was able to acquire data) $v'_{pupil}(t) \in \{valid, invalid\}$, $v'_{cr_i}(t) \in \{valid, invalid\}$ for the pupil and corneal reflections respectively.

The details of obtaining these four values will typically depend on the specifics of the particular eye-tracking employed, e.g. on details regarding illumination, camera characteristics, subject-camera-scene geometry, and hardware implementation. Also, for reference, the time t is typically sampled discretely, i.e. $t = t_0 + n\Delta_t$, where t_0 is the initial time, Δ_t is the time delta (the inverse of the camera frame rate), and n is the current time step.

2.3 Blink and Outlier Detection

The first stage in processing the incoming data stream is to detect blinks. There is little published literature regarding blink detection as related specifically to the input eye-tracking data streams with which we are working, though there are several studies examining blink detection in general video of faces (Bhaskar, Foo Tun Keat, Ranganath,

& Venkatesh, 2003; Crowley & Berard, 1997; Kawato & Tetsutani, 2004). Note that our purpose for detecting blinks is not necessarily to gain information regarding blinks per se, but to isolate regions of time where the incoming eye-tracking data stream is unstable or unusable. Blinks are thus temporally localized in order to track problems due to 1) the closing or opening of the eye-lids, and 2) the brief period of eye-tracking instability that precedes or follows lid aperture changes as the eye-tracking system attempts to regain pupil and corneal reflection coordinates. Blinks also serve as a natural boundary for demarcating the limits of saccades and fixations, however, and thus a distinction does need to be made between data regions that are unusable due to blinks and data regions that are unusable for other reasons, such as system tracking error. Figure 2.3 illustrates a typical blink sequence, showing how the image of the pupil is deformed and finally lost, together with the corneal reflections, as the eyelid closes.

The method that we describe here has been found in our work to be accurate in removing both blinks and data associated with periods of eye-tracking instability from our final processed data stream, and the parameters used are based on both efficacy and physiological properties (e.g. Caffier, Erdmann, & Ullsperger, 2003). We note first that the validity flag $v'_{pupil}(t)$ only reflects the presence of data and makes no claim as to the accuracy of that data. However, it is the case that every blink includes some amount of pupil data marked as *invalid*. Our algorithm thus proceeds by marking consecutive regions of *invalid* data as potential blinks. It then extends these regions based on a number of criteria that indicate system instability. Points that allow a marked potential blink region to expand into them are those which meet any of the following criteria:

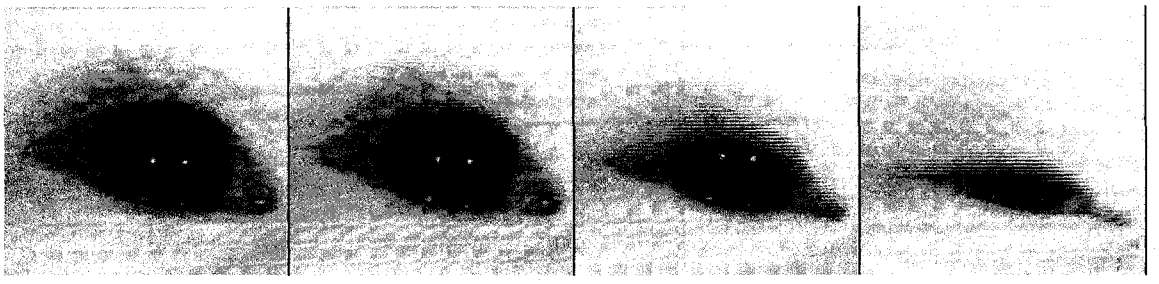


Figure 2.3: A blink sequence. As the lid close first the pupil deforms and then is lost, together with the corneal reflections. The image sequence is taken at 30Hz.

1) bad pupil aspect ratio, defined as:

$$\left| \frac{d'_{horizontal}(t)}{d'_{vertical}(t)} - 1 \right| \geq threshold_{aspect} \tag{2.1}$$

2) extreme relative changes along the vertical axis of the pupil, defined as:

$$2\left| \frac{d'_{vertical}(t \pm \Delta_t) - d'_{vertical}(t)}{d'_{vertical}(t) + d'_{vertical}(t \pm \Delta_t)} \right| \geq threshold_{vertical} \tag{2.2}$$

3) the absence of one or more corneal reflection points, i.e.

$$v'_{cr_i}(t) = invalid \text{ for any } i \tag{2.3}$$

After this process we have a set of expanded potential blinks. Expanded potential blink regions that are separated by less than $threshold_{t_\Delta}$ of non-blink data are then merged. The points which are merged over (i.e. the separating points of two merging expanded potential blinks) are added to a list of unusable, non-blink data points. Finally, all expanded potential blinks with durations lasting more than some threshold time $threshold_{t_{\min}}$ are added to a final set of valid blinks, and those that fail to meet the threshold time are added to the list of unusable, non-blink data points.

Next, outlier points are removed from the valid data stream and marked as unusable, non-blink data points. These outlier points meet any of the following criteria:

1) corneal reflections separated by extreme distances, i.e.

$$\left\| \boldsymbol{c}'_i(t) - \boldsymbol{c}'_j(t) \right\| > threshold_{\Delta_c}, \forall i, j, i \neq j \tag{2.4}$$

2) extreme distances between the pupil center and the centroid of the corneal reflections:

$$\left\| \boldsymbol{r}'(t) - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{c}'_i(t) \right\| \geq threshold_{\Delta_r} \tag{2.5}$$

3) being marked as invalid in the original pupil stream or any corneal reflection stream, i.e.

$$v'_{cr_i}(t) = invalid \text{ for any } i, \text{ or } v'_{pupil}(t) = invalid \tag{2.6}$$

The data that are neither blinks nor unusable non-blinks are the final set of constituent data that will build the referenced pupil coordinates:

$$\boldsymbol{r}(t) = \boldsymbol{r}'(t) - \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{c}'_i(t) \tag{2.7}$$

Since the corneal reflections and the pupil center will move together when the subject moves, this referencing to the average of the corneal reflections in Equation 2.7 results in a much greater resilience to subject motion. We will denote the set of times corresponding to blinks as $\mathbf{T}_{\mathbf{blinks}}$, the set of times corresponding to unusable, non-blinks $\mathbf{T}_{\mathbf{reject}}$, and the set of times corresponding to usable data as $\mathbf{T}_{\mathbf{valid}}$. Unless specifically noted, we will assume that the operations we discuss in subsequent sections refer to those data at times associated with $\mathbf{T}_{\mathbf{valid}}$. We will refer to pupil coordinates with the understanding that this refers to the normalized referenced pupil coordinates given by Equation 2.7.

2.4 Calibration

In order to convert pupil coordinates to the coordinates of the scene the subject is viewing, it is necessary to obtain reference points where both pupil and screen coordinates are known simultaneously. This process of calibration is typically accomplished by displaying to subjects objects with known positions on the screen. An example calibration routine is shown in Figure 2.4. Pupil coordinates which correspond exactly to a known calibration screen location are typically matched exactly; other locations in regions between calibrations are usually interpolated. There are at least two tradeoffs to consider in a calibration routine. First, the use of multiple calibrations targets increases the accuracy in locations far from the calibration targets, but comes at the expense of a longer time spent executing the calibration procedure. Second, the use of smaller calibration targets decreases the uncertainty as to where on the calibration target the subject is actually attending, but causes the calibration target to be less salient. Consideration of these tradeoffs is typically not important for the normal adult population, where multiple small cross-hairs can be used as targets, but is crucial for experiments involving young children and atypical subject populations, where a minimal number of larger targets, in conjunction with contingent sound and motion, would be more appropriate.

Though we have limited ourselves to the simpler methods of calibration in this thesis, we should note that there have been quite a number of recent developments which, in the future, might lead to even faster, more reliable calibration. For example, Morimoto, Amir, & Flickner (2002) have developed a technique in which it is possible to track gaze position without any user calibration via camera calibration, multiple light

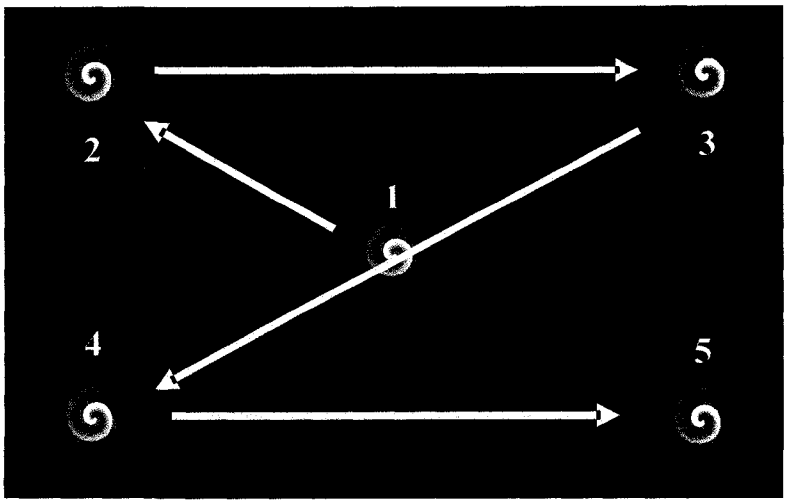


Figure 2.4: An example calibration routine. Calibration targets are shown, one at a time, in the locations indicated, in the order from 1 to 5, in order to elicit a gaze pattern in subjects which follows the white arrows, starting from position 1.

sources, and an assumed model of the subject's head. Similarly Yoo, Kim, Lee, & Chung (2002) use the cross ratios between the geometry of four corneal reflections and the pupil center for calibration-free eye-tracking. Both of these methods lead to high tracking error, unfortunately. More recently, Guestrin and Eizenman (2008) have reported a calibration technique that uses multiple corneal reflections in conjunction with multiple eye cameras and known geometry in order to reduce the calibration phase to a single calibration target with high accuracy.

2.4.1 Bootstrapping

In order to match the screen coordinates of a calibration target to a pupil position, we must locate the periods of time where the pupil has fixated the calibration target. To accomplish this, we must segregate the fast transitory motion to the target (white arrows in Figure 2.4) from the stable process of viewing the target. This is a fixation

identification problem (Section 2.5). However, unlike the standard fixation identification problem, there is no knowledge of the scale of eye-movements. In other words, we can't calibrate without finding fixations, but we can't find fixations without calibrating. This situation can be alleviated by assuming a reasonable default coordinate system or default operating scale and performing fixation identification under those defaults in order to locate pupil locations corresponding to calibration targets. This can be done by 1) assuming that the distribution of saccades and fixations in the data follows some natural distribution over all subjects, e.g. assume that 20% of all pupil data is from saccades and the remainder is from fixations and using this information to obtain parameters for each subject; or 2) performing an exploratory analysis and obtaining a reasonable estimate which is broadly applicable to all subjects. In homogenous subject populations, such as typical adults, both methods would likely perform well, but the first method would help compensate for changes in the operational characteristics of the eye-tracking system or the experimental setup. In heterogeneous subject populations, however, assuming a distribution for each individual could lead to greater system instabilities and the second method, assuming a global default for all subjects, would be preferred, though care would have to be taken to ensure reproducible testing conditions. It is important to remember that the fixation identification at this stage is only for the purposes of bootstrapping calibration, and does not reflect the final identification of saccades and fixations. For our work we have chosen a bootstrapping method using velocity threshold in conjunction with hysteresis (Section 2.5.2).

2.4.2 Target matching

We want to find the fixation which carries the most information regarding the pupil coordinates associated with the calibration target. Every calibration target c is displayed over a range of time $\mathbf{T}_c = \{t_0^c, ..., t_n^c\}$ and every fixation f identified by bootstrapping in Section 2.4.1 has an associated set of times $\mathbf{T}_f$. Note that the set of times associated with the calibration target, $\mathbf{T}_c$, does not perfectly correspond to the set of potential fixations because there is a lag as the eye responds to the onset of a new target. Consequently an adjusted calibration time set $\mathbf{T}_{c^\Delta} = \{t_0^{c^\Delta}, ..., t_n^{c^\Delta}\}$ associated with c is defined, with $t_i^{c^\Delta} = t_i^c + \Delta \quad \forall t_i^c \in \mathbf{T}_c$. Every fixation f' such that $\mathbf{T}_{f'} \bigcap \mathbf{T}_{c^\Delta} \neq \emptyset$ is a candidate for being the representative fixation for the calibration target c. We choose the longest fixation belonging to the set of candidate fixations $\mathbf{F'}$ as the representative fixation f_c:

$$f_c = \mathbf{arg\,max}_{f' \in \mathbf{F'}} \left\| \mathbf{T}_{f'} \right\| \tag{2.8}$$

Note that the distance measure $\|\bullet\|$ in Equation 2.8 can correspond either to the cardinality of $\mathbf{T}$, i.e. the number of constituent valid time points, or the range of $\mathbf{T}$, i.e. the maximum element minus the minimum element in $\mathbf{T}$.

From f_c we can extract representative pupil coordinates, $\boldsymbol{r}_c$, from the set of reference pupil coordinates $\mathbf{R}_c = \bigcup_{t \in \mathbf{T}_{fc}} \boldsymbol{r}(t)$. This extraction can occur by taking some variant of the mean or median on the positions in $\mathbf{R}_c$. The decision regarding the function used for the extraction process will depend on particular noise and stability concerns. We have found that taking the median for each coordinate axis in $\mathbf{R}_c$ is simple

and works well. This strategy is particularly effective against trailing or leading outlier points.

A final validation step rejects or accepts each calibration fixation f_c, and, by consequence, the entirety of c, based on two quality criteria. First, we reject calibration fixations if they contain too few points or extend over too little time to be reliable, i.e. f_c is rejected if $\|\mathbf{T}_{f_c}\| < threshold_{t_{calmin}}$. Second, we reject calibration fixations that are too far spatially (based on some predetermined criteria) from the mean or median of all calibration fixations in that position. In other words, if we let c_i^p be the ith calibration target associated with calibration position p (e.g. in Figure 2.4, $p \in \{1..5\}$), and let $f_{c_i^p}$ and $\boldsymbol{r}_{c_i^p}$ be the associated reference fixation and pupil coordinates for that calibration, respectively, we reject $f_{c_i^p}$ if $\boldsymbol{r}_{c_i^p}$ is too far from the mean or median pupil coordinates, $\boldsymbol{r}_{c_\mu^p}$, of all other fixations associated with that same calibration point, i.e. $f_{c_i^p}$ is rejected if $\left\|\boldsymbol{r}_{c_i^p} - \boldsymbol{r}_{c_\mu^p}\right\| \geq threshold_{d_{calmax}}$.

2.4.3 Translating pupil coordinates

Once calibrations have been established, it is possible to convert valid pupil coordinates anywhere into the screen coordinates of the display shown to the subjects. This process of interpolation can take many forms, and here we describe a simple strategy based on piecewise linear interpolation on a set of 5 calibrations (Figure 2.5), as is commonly used in situations where minimizing calibration time is necessary due to problems with attention in subjects (e.g. as would be the case when working with children).

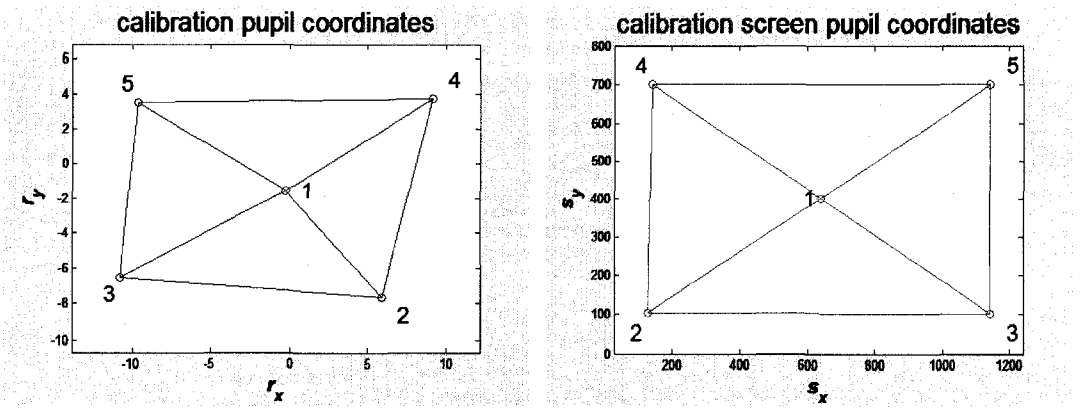


Figure 2.5: *left:* Calibration pupil coordinates of 5 calibration locations in device-dependent units; *right:* screen coordinates for calibration locations in pixels.

Every valid pupil coordinate must have some translation into screen coordinates. A simple, but effective conversion process is to first determine into which triangle of the calibration pupil coordinates the unknown target pupil location falls (i.e. any unique set of three calibration points that includes the calibration at position 1). This is done simply by determining the angle of the target pupil coordinate in relation to the calibration pupil coordinates with the center calibration (Calibration 1 in Figure 2.5) as the zero coordinate. Once the calibration region is established, a simple linear interpolation based upon the triangle spanned by the region can map the pupil coordinate to a screen coordinate.

2.4.4 Recalibration

Over the course of a long experiment it is necessary to periodically recalibrate the system in order to maintain accurate localization of the subject's gaze. Calibration drift can be caused by a number of factors but is typically caused by either conscious or unconscious subject motion, e.g. the subject jerking his head to the side at the appearance of some

novel stimulus. Since any fixations where pupil and screen coordinates are known simultaneously can be used to help calibrate the eye-tracking system, it is advantageous to consider time-varying strategies for calibration. For example, Hornof and Halverson (2002) draw a distinction between explicitly required fixation locations for calibration and implicit targets. These implicit targets can be used to gauge the error in the system and can be embedded so as to not disrupt the flow of an experiment. As another example, if a central target is used to attract the attention of subjects at the beginning of each trial on a multi-trial experiment, this information can be used to recalibrate the system. There are several strategies for recalibration:

1) *no recalibration* – The standard method in eye-tracking experiments is to use the initial set of calibrations throughout the course of the experiment. This has the advantage of requiring no post-hoc reanalysis and thus is very applicable to real-time eye-tracking applications.
2) *static calibration* – When reanalysis is possible, a better solution is to use all available calibration data to build a master set of pupil-screen calibration mappings, i.e. to build a set of canonical calibrations, such that the static set of calibrations is most representative of the experiment as a whole. Building this set of canonical calibrations falls naturally out of the calibration rejection by distance heuristic found at the end of Section 2.4.2.
3) *temporal interpolation* – Given that calibration targets are periodically represented to subjects, temporal interpolation acts by viewing the pupil coordinates associated with a calibration in a particular position in a temporally-

dependant fashion. The simplest technique is piece-wise linear interpolation of pupil coordinates between calibration times, however more advanced techniques could use smoother, higher-order surfaces.

4) *nearest neighbor* – temporal interpolation typically views the changes in calibrations as smooth and continuous. This is reasonable when calibration changes are due to the accumulation of slight movements, such as when a subject slowly slides down in his chair as he tires. However, in many cases changes in calibration are due to fast, abrupt changes. In these situations it makes sense to view the calibrations as "all or nothing" rather than slowly changing. The *nearest-neighbor* method for recalibration simply uses the calibrations that are temporally closest to the trial at hand as the representative calibration. This method preserves the integrity of each trial while adapting to changes in calibrations as the result of system or subject changes.

5) *manual recalibration* – Another popular recalibration technique is to manually restart the calibration procedure whenever the experimenter notices motion in the subject or finds the current set of calibrations wanting, either through observation or through heuristics (such as checking for the deviance on centering stimuli). Of course, this method relies on the experience of the experimenter, and though automated procedures should be possible, they have not yet come into widespread use. We should note that though it is typically assumed that, when *manual recalibration* is used, that it is the only recalibration method available, the generated mappings can be combined with the previous methods we have introduced.

The choice of a recalibration technique may seem like a somewhat minor detail. However, there are considerable differences that can be caused by choosing one technique for another, sometimes dramatically altering observed patterns of results (Figure 2.6). As with all choices in experimental analysis, the consideration of which technique is most appropriate depends heavily on the specifics of the subjects and experiments that are to be applied. In compliant typical adults, for example, a single calibration at the beginning of the experiment might suffice, especially when combined with a bite bar or other means of constraining the head. For a difficult population, such as young children with developmental disabilities, subject motion, inattention, or problems with affect can lead to a staggering lost of data if the proper techniques for calibration are not applied. In our experience, periodic automated recalibration with the nearest neighbor technique, with occasional manual recalibration under certain circumstances, suffices for obtaining high-quality data for subsequent analysis. However, measures still need to be employed to evaluate the accuracy of the eye-tracking data.

2.4.5 Gauging error

Since calibration is our only means of mapping pupil coordinates to screen coordinates, it is difficult to gauge true error, as the subject could choose to fixate on, for example, the edge of the screen rather than the calibration, and do so in such a consistent manner as to be indistinguishable from gaze directed towards the calibration target. However, for practical purposes, it is usually assumed that this is not the case, and a typical method for

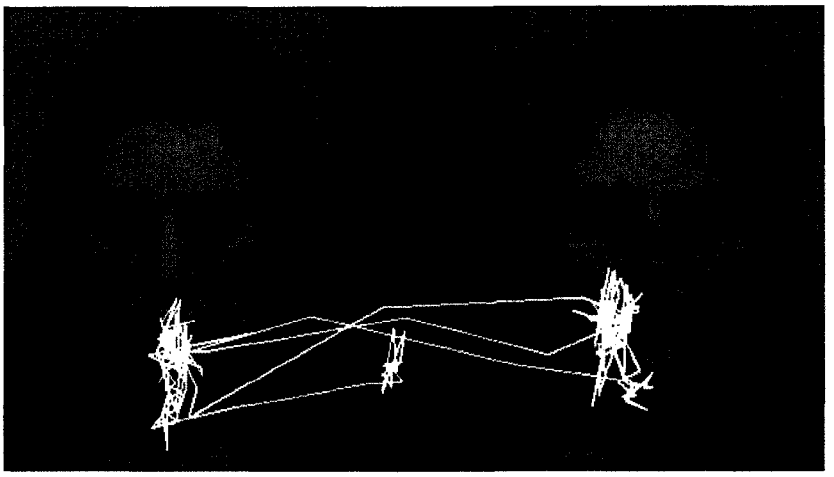


Figure 2.6: Example of differences caused by choices in recalibration techniques. The top two scan paths are the nearest neighbor recalibration (green) and static calibration (blue) whereas the bottom uses no recalibration (yellow). The subject is a 17 month old female with language delay.

gauging error is to hold out a set of calibrations for measuring the error of the experiment. For example, a calibrated system will be subject to any number of evaluation targets centered at various locations on the screen. The difference between the calibration-calculated position and the evaluation target position is used as a measure of the tracking error. This leads us to a question: what positions should be evaluated? Evaluating only the center of the screen is a biased approach, since it is expected that the system is best calibrated at this position. Examining all calibration points every trial is time-consuming and could be considered excessive. In addition, the use of evaluation targets at all seems somewhat wasteful, since these positions could themselves be used for recalibrating the pupil-to-screen mapping. Furthermore, evaluation points only give error regarding the locations that are evaluated; they do not provide information about the practical effect on the scanpaths. For example, the error at the upper left corner of the

screen does not matter as much as the error in the lower right corner of the screen when the stimulus is being shown in the lower right corner. Finally, in more difficult subject populations, it is not always the case that subjects will attend to the evaluation target. In many cases, the trial contains data but the single evaluation measure fails.

An alternative is to use the differences between the various calibration techniques as a measure of error. For example, we use the mean difference in trajectories calculated by the nearest neighbor recalibration and static calibration as an error measure. Since the static calibration, by definition, is the best single set of calibrations for the experiment as a whole, it can be viewed as the baseline. Since the nearest neighbor recalibration method is an adaptive technique, the difference between nearest neighbor recalibration and static calibration can be seen as reflecting changes in the subject or eye-tracking system away from this baseline. If this difference is large, it is likely that the current set of trials represent too large of a deviation from the baseline to be trustworthy. This method has the advantage of reflecting the impact of a calibration change on the entire scanpath under consideration rather than the impact at select few points. Furthermore, this method also reduces data loss as it allows trials to be retained where changes to calibrations are inconsequential to the resultant scanning pattern, and allows us to use all available information for recalibration when possible.

There are of course, different variations that could be applied on the differences of techniques method. For example, rather than viewing the static calibration as a baseline, an assumption which might in fact be too strict, it should be possible to use a sliding window of calibrations as the baseline instead. In addition, taking the maximum difference between the trajectories rather than the mean difference could provide a

stricter measure for the trajectory deviance. It is also important to note that error measurements are heavily impacted by previous steps in the data preparation pipeline, and that the identification of true calibration fixations is critical to the validity of any recalibration technique.

2.5 Fixation Identification

After the system has been calibrated, gaze patterns are typically dissected through fixation identification algorithms (Duchowski, 2003; Salvucci and Goldberg, 2000). These algorithms take as an input the raw stream of converted scene coordinates and group the data points of that stream into a series of saccades (rapid, ballistic movements of the eye) and fixations (periods where the point of regard by the eye is spatially relatively stable). This dichotomous parsing is employed for two reasons. First, there is psychological and neurophysiological evidence that visual field processing is suppressed via saccadic masking during rapid movements of the eye (Burr, Morrone, and Ross, 1994; Erdmann and Dodge, 1898) and so it makes sense to discard saccades from experiments that focus on conscious perception. Second, the ability to deal in quanta of fixations simplifies analysis and interpretation, as each fixation can be seen as being associated uniquely with a particular spatiotemporal location which in turn can be associated with particular perceptual qualities of the visual scene. Fixations can then be aggregated at many different levels, resulting in a wealth of psychophysical measures such as the total amount of time spent in fixations, the average duration of fixations, the number of fixations, latency of the first fixation after stimulus presentation, etc. (Inhoff

and Radach, 1998; Jacob and Karn, 2003). Thus finding fixations serves many useful purposes, from aiding in interpretation, to data reduction, and finally to being objects for interpretation in their own right.

The most common fixation identification algorithms operate in a greedy fashion (Salvucci & Goldberg, 2000). Typically, these algorithms begin with the assumption that fixations beneath a certain amount of time t_{min} are too short to be considered physiologically realistic due to the inherent latency necessary for the preparation of a saccade (Leigh & Zee, 2006). The algorithms then impose other constraints on fixations, such as spatial constraints on the relationships between all screen points in a fixation (dispersion-based threshold algorithms) or the point-to-point velocities in a fixation (velocity-based threshold algorithms). Additional constraints enforce consistency with other steps in the processing pipeline, such as a constraint that no times $\mathbf{T}_f$ associated with a fixation f overlap with $\mathbf{T}_{\mathbf{blinks}}$, i.e. with any times where blinks are recorded. The greedy fixation begins at the start of a screen position data stream to be analyzed, selecting as a candidate fixation a consecutive block of time where all constraints are satisfied. It then expands this candidate fixation until spatial constraints are violated, adds the constraint unviolated portion of the candidate fixation to a list of fixations, and then reinitializes the process for the next candidate fixation on next non-overlapping consecutive block of time where all constraints are satisfied.

The greedy nature of these fixation algorithms makes them extremely efficient, allowing them to keep track of a minimal state (i.e. the current fixation in the making) which makes them well suited for real-time applications. However, though greedy

algorithms perform well, for post-hoc analyses there are other algorithms which might perform even better by leveraging not only prior history regarding the scanpath but also future information. For example Goldberg and Schryver (1993) use a minimum-spanning tree method, Privitera and Stark (2000) use k-means clustering, and more recently, Santella & Decarlo (2004) have used a mean-shift procedure and Urruty, Lew, Djeraba, and Simovici (2007) a projection clustering technique. Use of these techniques is not widespread. It is much more common for off-the shelf eye-tracking systems to contain simple greedy algorithms and it remains to be seen whether these modern approaches will find widespread acceptance and use. In Sections 2.5.1 and 2.5.2 we discuss some of these simple algorithms; in Chapter 3 we will compare some of these methods under a new framework.

2.5.1 Dispersion-based Methods

Dispersion-based threshold algorithms mark a segment of consecutive points in the scanpath as a fixation if those points obey certain temporal and spatial constraints. The temporal constraint is a duration requirement: if the duration of a fixation is less than a threshold time t_{min}, it is judged to be non-physiological and is marked as invalid. The spatial constraints are more variable and will be discussed for several algorithms.

Here we consider four variations: (1) a pure distance dispersion algorithm; (2) the centroid distance scheme by Anliker (Anliker, 1976; Duchowski, 2003); (3) the position variance-method, also due to Anliker (1976); and (4) Salvucci's I-DT algorithm (Salvucci & Goldberg, 2000).

1) *Distance Dispersion Algorithm* – For a fixation to be valid under the distance dispersion algorithm, each point in that fixation must be no further than some threshold d_{max} from every other point. This is perhaps the most intuitive measure, but is less popular than other simple dispersion algorithms because every fixation point must be checked against every other fixation point, resulting in $O(n^2)$ operations.

2) *Centroid-Distance Method* – Anliker's centroid-distance method (Anliker, 1976) requires that M of N points be no further than some threshold c_{max} from the centroid of the N points. Since the M of N criteria has denoising properties, for comparability in our studies, we have set M=N. Also, it is possible to use either a consistent version of this algorithm, where, whenever the fixation is being expanded (Salvucci & Goldberg, 2000; Widdel, 1984), we recompute the distance of all points in the fixation to the centroid, or a fast version, where we only check the distance of the new point to be added. Here, we use the more consistent version.

3) *Position-Variance Method* – This method (Anliker, 1976) is a variant of the centroid-distance restricted algorithm where it is required that M of N points have a standard deviation of distance from the centroid not exceeding σ_{max}. Again we set M=N and use the consistent interpretation of the algorithm.

4) *Salvucci I-DT Algorithm* – Salvucci's fixation identification by dispersion threshold algorithm (Salvucci & Goldberg, 2000) requires that the maximal horizontal distance plus the maximal vertical distance is less than some threshold m_{max}. This algorithm is a fast approximation to the distance dispersion algorithm,

as only the extreme horizontal and vertical positions need to be compared against when adding a new point to a fixation.

2.5.2 Velocity-based Methods

Velocity-based threshold algorithms detect saccades rather than fixations directly. Typically a specific velocity, v_{max}, is chosen as a threshold (Anliker, 1976; Duchowski, 2003). Points that exceed this velocity are considered saccadic movements; points below this are potential areas of fixations. Thus the spatial constraint in velocity-based threshold methods is a velocity constraint. The velocity can be determined in any number of ways, for example by a simple point-to-point difference divided by the sampling time, or through various combinations of filtering and prediction. Typically, the angle of the scan trajectory is not taken into account when considering instantaneously measured velocity (e.g. two consecutive segments on the trajectory with a roughly equivalent angle could be considered part of the same motion), though incorporation of this information could increase the accuracy of velocity-methods.

Though a minimum time constraint, t_{min}, is not commonly used in velocity-threshold algorithms, it can naturally be incorporated in one of two ways. First, the use of a minimum time requirement for fixation duration could be used in the same way as in dispersion algorithms (e.g. as an constraint on the initial candidate pool of fixation points in a sliding window). Second, and more common, all fixations detected with this method can be rejected outright if they do not meet the time requirement.

Some simple variations for controlling noise in velocity threshold algorithms also exist. For example, one can build a hysteresis into the saccade detection controller,

marking a segment as a saccade if the detected velocity exceeds some higher threshold $threshold_{high}$ and marking the end of the saccade only when the velocity falls below some lower threshold $threshold_{low}$.

2.6 Discussion

It becomes immediately apparent that there is not necessarily one perfect processing pipeline for all applications. Indeed, the particular choices in data processing for eye-tracking will hinge on the demands required by the eye-tracking users. In cases where only qualitative information is sought, it may suffice to skip fixation identification altogether. In cases where real-time operation and response are required, it may make sense to use the most simple, parsimonious algorithms available. For post-hoc analyses the tradeoffs are even more complex, as there is no one agreed upon best method for performing eye-tracking analysis. In the case of post-hoc analysis one must consider the tradeoff between the robustness and accuracy of more advanced and complex methods versus the likelihood that others might actually adopt these techniques and thus be able to reproduce one's findings or generate results which can be compared.

Many of the techniques presented in this chapter belong to the repertoire of standard eye-tracking methodology. However, it is also the case that this methodology relies on a number of assumptions, some of which have been taken far out of context in regards to their applicability. For example, the use of automated algorithms for discriminating between fixations and saccades originally found great popularity in studies of reading (Rayner, 1998). The parameters necessary for fixation identification in

reading are thus well-studied and well-known. However, nowhere can it be found that the parameters from reading are applicable to, for example, search tasks. In the following chapter we will take a closer look at the problem of parameters in fixation identification. We will show that it does not make sense to consider an optimal set of parameters for fixation identification, that the traditional delineation between fixation and saccade is an arbitrary one, and that characterizing the behavior of gaze patterns as a distribution provides a more powerful and complete description.

2.7 Chapter Summary

- We have presented the details of the eye-tracking processing pipeline from the initial stage of acquiring the image of the eye to the delineation of saccades from fixations. Specifically we have covered:
    - Detection of corneal reflections and the center of the pupil
    - Detection of blinks and erroneous data
    - Calibration for converting pupil coordinates to stimulus coordinates
    - Algorithms for separating saccades from fixations
- We have discussed some issues that are usually not addressed in standard texts on eye-tracking methodology, particularly:
    - The need for bootstrapping in calibration, i.e. that an initial pass must be conducted to locate target fixations before those target fixations can be used as markers for calibration
    - Different techniques for recalibration of data for post hoc analysis

    - Novel practical methods for gauging eye-tracking calibration error
- We have pointed out that many of the algorithms that go into the eye-tracking pipeline rely on heuristics that have not been wholly validated and suggest that the impact of these assumptions on interpretation can be quite severe.

Chapter 3

Distributional Modeling of Fixations and Saccades

In this chapter we will examine the effects of parameter choices on one particular fixation measurement, the mean fixation duration, as this measure has been correlated with increasing cognitive load (Fitts, Jones, & Milton, 1950; Goldberg & Schryver, 1993; Jacob & Karn, 2003; Crosby, Iding, & Chin, 2001), and thus represents a measure that has been taken to have a direct psychological analogue (Section 3.1). In traditional analysis, one set of parameters would be chosen in order to characterize saccades versus fixations. For example, using a spatial parameter of 1° and a minimum time duration of 100 ms, we might find that typical individuals spend on average 35 ms longer per fixation looking at faces than individuals with autism, and we would thereby conclude that typical individuals exhibit a greater cognitive load when viewing faces. It is important to note that the decision to choose those particular sets of parameters would be quite a subjective decision, as the parameters would have been selected in the absence of ground truth (e.g. neuronal recordings indicating a saccade). Thus, fixation identification algorithms in eye-tracking analysis serve mainly a descriptive purpose, or provide other analytical conveniences, such as data reduction.

In contrast to previous work, we will not presuppose that there exists an optimal choice for parameter selection. We will instead cover a large swath of parameters and algorithms and examine the systematic and interpretive differences that arise as a result of this manipulation (Section 3.2). We will find that by changing our parameters we can nullify or even negate observed trend, implying that how the popular method of choosing

a single set of parameters and a single algorithm is incomplete when considering the differences between the scanning of different classes of objects. We will demonstrate, however, that our straightforward approach leads to a deeper understanding of the space of parameter variation, and present a simple linear interpolation model (SLIM) which characterizes the effects of parameter changes at all parameter choices. With this understanding, we will then use a variation of SLIM to examine the scanning patterns of children with autism spectrum disorder, developmental delay without autistic symptoms, and typically-developing individuals (Section 3.3). We will show how examining the parameters of SLIM provides us with evidence consistent with known face processing abnormalities in autism. We will also demonstrate that the difficulties with using a single parameter set also apply to comparisons of different subject populations as well. We will continue by showing how the mean fixation duration is potentially related to another common measure, the number of fixations, and how a deeper knowledge of the distributional aspects of the number of fixations may suggest a fractal structure to gaze in free search, furthering strengthening our case against a single choice of algorithm and parameters for eye-tracking analyses (Section 3.4). We will conclude by discussing how the parameter problem in fixation identification, the simple linear interpolation model (SLIM), and the fractal qualities of scanning all are reflections of cognitive factors influence gaze strategy and discuss the nature of these factors and how they may be eventually uncovered (Section 3.5).

3.1 The Parameter Problem in Fixation Identification

For a calibration technique or a feature-identification algorithm it is possible to either create benchmarks or use existing benchmarks to calculate the error. This is not the case with fixation and saccade identification. The discrimination between saccades and fixations is clear from neurophysiological perspective, where certain populations of neurons are firing or not (Leigh & Zee, 2006), but from the perspective of a post-hoc analysis of the scanpaths themselves, the boundaries between the two can be much more ambiguous. For example in Figure 3.1 we can see that, by varying spatial parameters of the distance dispersion algorithm, we can obtain several differe7nt interpretations of what constitutes a fixation.

The variability caused by changes in the parameters of fixation identification algorithms and the difficulty in comparing different algorithms has been long noted. For example, Karsh and Breitenbach (1983) commented on the extensive qualitative changes that could occur when fixation parameters were varied systematically. Similarly, by parameter variation of a fixation identification strategy for a set of subjects sequentially fixating a grid of dots, Widdel (1984) showed that considerable differences could be generated in the location and number of reported fixations and saccades. The general conclusion from these studies was that every analysis using eye-tracking in conjunction with fixation identification should report the exact algorithm and parameters used. However, since every study using eye-tracking is different, it was also generally conceded that the choices in analyses be made for the situation at hand. This practical approach has been the unwritten rule of eye-tracking research for many years, and despite the fact that the first fixation identification algorithms arrived more than thirty years ago

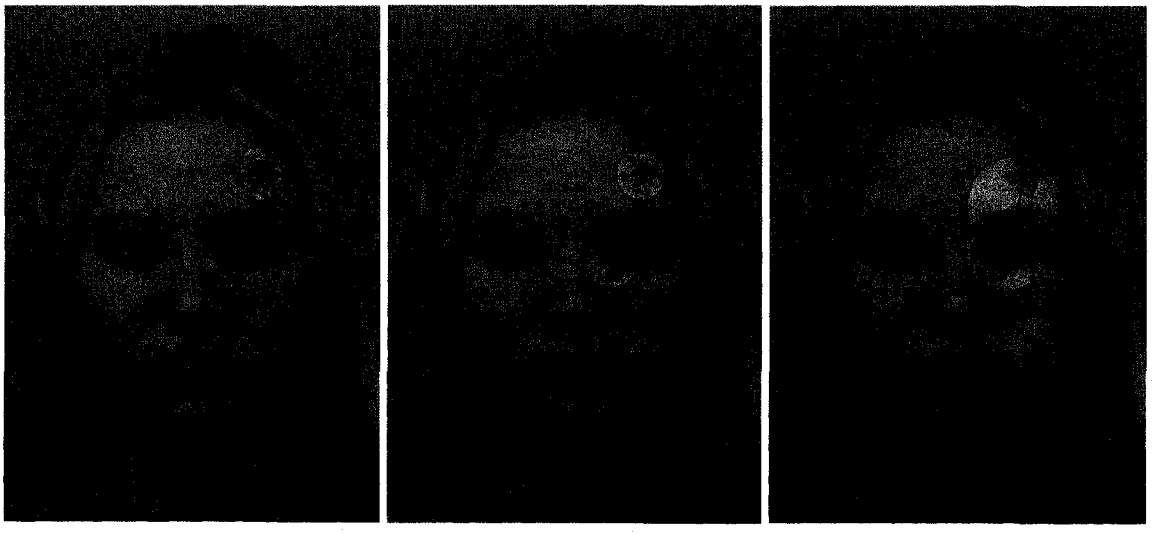


Figure 3.1: Changes in identified fixations caused by varying spatial parameters. Fixations are identified with the distance dispersion algorithm with a maximal dispersion of 0.9° (left), 1.8° (center), and 3° (right). Yellow circles are the computed fixations and green lines are the scanpath. The subject is a 2 year old male child with ASD viewing a static image of a face.

(e.g. see Anliker (1976)), and the fact that most researchers who use eye-tracking are well aware of its limitations, the use of a single fixation identification algorithm with a single set of parameter choices has persisted as the dominant means of analysis. However, though these recommendations may be practical, they do not address the issue of how exactly parameters should be chosen in the first place. Taken to an extreme, these recommendations are a license for the first explorers of a field to arbitrarily pick parameters. Consequently, today, not only is there no consensus as to which fixation identification algorithm should be used, there is no agreement, even within a particular algorithm or a particular task, as to what choice of parameters should be employed.

The use of a single set of parameters and algorithm for fixation identification is also problematic for at least three other reasons. First, it means that studies which use different methodologies for fixation identification may never be directly comparable, though their results and interpretations may have broad implications. It becomes all too easy, also, to cite as support a study which is in agreement with one's results, and brush away other work that uses different parameters because they use different parameters. Second, aligning researchers to the same algorithms and parameters even within a single field is a monumental task, one which would require not only the agreement of the researchers themselves, but the cooperation and assistance of the growing list of eye-tracking manufacturers and writers of eye-tracking analysis software. Third, it assumes that the choice of parameters is reasonable in the first place, and will lead to interpretable results. This puts the onus of discovery on the first researcher who designs an experiment; it further constrains subsequent researchers who might be interested in slightly different methods. In many cases, the choice of parameters is made on subjective determination of whether a set of fixations appears reasonable and so an investigation into the effects of different parameters is certainly understandable.

To complicate matters further, there is a long list of possible fixation measures that one could consider when performing analysis. Given that a particular definition of a fixation is reasonable, one could consider: the total amount of time spent in fixations, the number of fixations, the mean fixation duration, the frequency of fixations, the latency to reach the first fixation, the duration of that first fixation, and so forth (Jacob & Karn, 2003; Radach & Kennedy, 2004). Similar measures could be applied for saccades. There have been very few systematic analyses of these measures, and it is not clear how

the measures themselves are impacted by choices in fixation identification algorithms. In many ways, the lack of fundamental research into these measures has held deep research in eye-tracking back, as it has left many eye-tracking researchers with a quandary: does one 1) peg one's work to the analytical design and choices of others, preferring not to investigate the possibly great variation which could arise from these decisions in design?; or 2) choose methods that are sufficient for the task at hand, based on subjective criteria?; or 3) report the results associated with the best set of parameters and measures one can find?

Faced by a confusion of algorithms and parameters, we might be tempted to throw our hands up in surrender and either decide that measures on fixations have no inherent value or decide to escape into a corner of the parameter space with neurophysiological constraints, such as the diameter size of the foveola, as our shelter. However, though both of these decisions would not be completely without merit, we believe that there is in fact a better solution, one which will allow us not only to model and interpret the reversals that we have observed, but will also give us a hope for unifying the disparate results of prior work which to date defy comparison for lack of a common language.

3.2 The Simple Linear Interpolation Model (SLIM) for Mean Fixation Duration

Karsh and Breitenbach (1983) remarked upon the "amorphous fixation measure", i.e. how qualities of the scanpath changed and varied as parameters of the fixation

identification algorithm changed. In this section, we will show how the current use of fixation identification algorithms is not necessarily amorphous, but rather, incomplete. We will accomplish this by performing an experiment by which parameter choices and algorithms used in scanpath analysis are systematically varied. From this, we will show that the dependence of mean fixation duration on parameters can be well expressed in terms of a small number of coefficients, and that these coefficients give us a more complete picture of the actual scanpath dynamics. This section is based off work originally presented in (Shic, Chawarska, & Scassellati, 2008a).

3.2.1 Fitting a Plane through the Origin

15 typically developing toddlers (mean age 27 months; range 18 to 33 months) were shown 6 color images of faces (stimuli derived from Lundqvist, Flykt, & Ohman, 1998) and 6 block designs (Figure 3.2) at a distance of 75 cm on a 24" (61 cm) widescreen monitor (16:9 aspect ratio). Each image, including the grey background, was 12.8° x 17.6°. Eye-tracking data were obtained simultaneously with a SensoMotoric Instruments IView X RED table-mounted 60Hz eye-tracker. Stimulus images were preceded by a central fixation to refocus the child's attention and were then displayed as long as was required for the child to attend to the image for a total of 10 full seconds. Actual trials could last longer than 10 seconds; however, to maintain comparability, only the first 10 seconds of each trial were used in our analysis, and trials which did not contain at least 5 seconds of valid eye-tracking data, or which did not meet automated quality criteria (as discussed in Section 2.4.5), were discarded. Furthermore, only data falling within the stimulus image area were considered in analysis. This task was

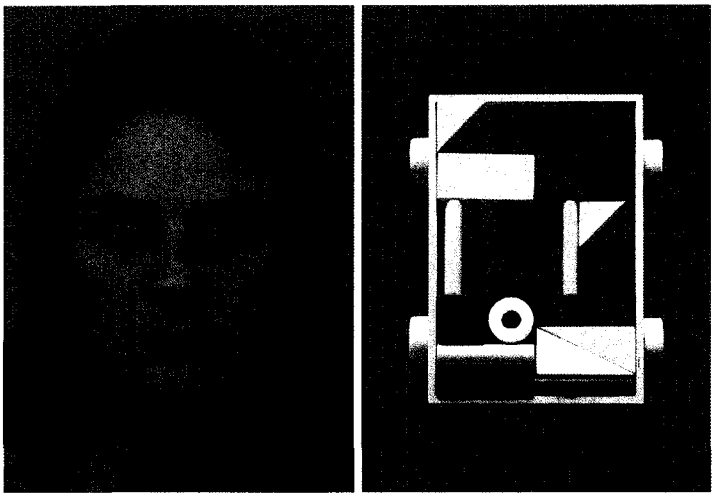


Figure 3.2: Examples of Experimental Stimuli used in this study. *Left:* faces from (Lundqvist et al., 1998). *Right:* abstract block patterns taken from a children's toy.

embedded within a visual paired comparisons recognition task (Fantz, 1964), i.e. was followed by exposure to both the same face and a novel face on either side of the screen. We do not consider the recognition phase here, but do in subsequent work (Chawarska & Shic). A total of 41 trials on blocks and 27 trials on faces were obtained. Loss of data was typically caused by poor affect (e.g. crying) or poor attention and was within the range expected for this subject population. Results were aggregated at the level of a trial (i.e. mean fixation times reported are means of trial means).

Five different algorithms from Section 2.5 were analyzed: four dispersion algorithms (the distance method, centroid method, variance method, and I-DT algorithm) and one velocity algorithm without hysteresis. In order to characterize the behavior of the algorithms over physiologically reasonable parameter settings, we examined mean fixation duration (defined as the amount of time spent in fixations divided by the total number of fixations) for each algorithm under a grid of uniformly sampled temporal and

spatial constraints. The temporal constraint was the minimum duration requirement ($t_{min} \in$ [50ms,250ms], N=13). The spatial constraints were computed by matching the mean fixation time range of each algorithm to the range of the distance dispersion algorithm over ($d_{max} \in [0.6°,5.1°]$, N=16), giving us a spatial parameter **s** ranging from a minimum value $\mathbf{s_{min}}$ to a maximum value $\mathbf{s_{max}}$ (i.e. $\mathbf{s} \in [\mathbf{s_{min}},\mathbf{s_{max}}]$) for each algorithm. A multiple linear regression (without offset) was applied to the mean fixation times t_{fix} as a function of t_{min} and **s** in order to find the temporal and spatial slopes of each algorithm ($\mathbf{slope_t}$ and $\mathbf{slope_s}$, respectively):

$$t_{fix}(t_{min},s) = \text{slope}_t\, t_{min} + \text{slope}_s\, s \qquad (3.1)$$

As shown by Table 3.1 (see also Figure 3.3), over a minimum time duration from 50 ms to 250 ms and a spatial extent ranging from 0.6° to 5.1° (distance algorithm equivalent), mean fixation duration for all algorithms is essentially a linear function of parameters. The exact characterization of the linear effect differs, however, between algorithms and between stimuli.

The difference in slopes between stimuli classes is caused by changes in the scanning distribution as a result of changing image properties. We note that for faces, in comparison to blocks, the spatial slopes are lower and the temporal slopes are higher. As we will show later, higher spatial slopes correspond to denser scanning patterns and the results for the temporal parameter should be related and correlated with changes in spatial parameters. We should note that though the linear relationship shown here is simple, it is not necessarily obvious. Doubling the spatial parameter of the distance dispersion

Faces

Method	$slope_t$	S	s_{min}	s_{max}	$slope_s$	R^2
Distance	0.98	d_{max}	0.6°	5.1°	151	.996
Centroid	1.10	c_{max}	0.4°	3.4°	217	.995
Variance	1.20	σ_{max}	0.15°	0.85°	893	.992
I-DT	0.85	m_{max}	1.5°	8°	102	.998
Velocity	0.58	v_{max}	18°/s	81°/s	9.47	.993

Blocks

Method	$slope_t$	S	s_{min}	s_{max}	$slope_s$	R^2
Distance	0.75	d_{max}	0.6°	5.1°	178	1.0
Centroid	0.89	c_{max}	0.4°	3.6°	258	.999
Variance	0.81	σ_{max}	0.15°	0.75°	1224	.999
I-DT	0.57	m_{max}	1.5°	8°	118	1.0
Velocity	0.51	v_{max}	18°/s	81°/s	9.69	.988

Table 3.1: Mean Fixation Times of Algorithms as a Linear Function of Parameters for Faces (top) and Blocks (bottom). Note that the R^2 reported are affected by traditional error estimations on regressions without an offset term (see Gordon (1981) and Section 3.3.3 Limitations and Implications).

algorithm quadruples the area. If the distribution of points was linear, we would expect a quadratic effect on mean fixation time; if it was diffusive, we would expect an effect for some characteristic length. Instead, we see a simple linear relationship. We will return to resolve this issue in Section 3.4.

The change in slopes for different algorithms follows from the different assumptions each algorithm makes regarding spatiotemporal scanpath behavior. For

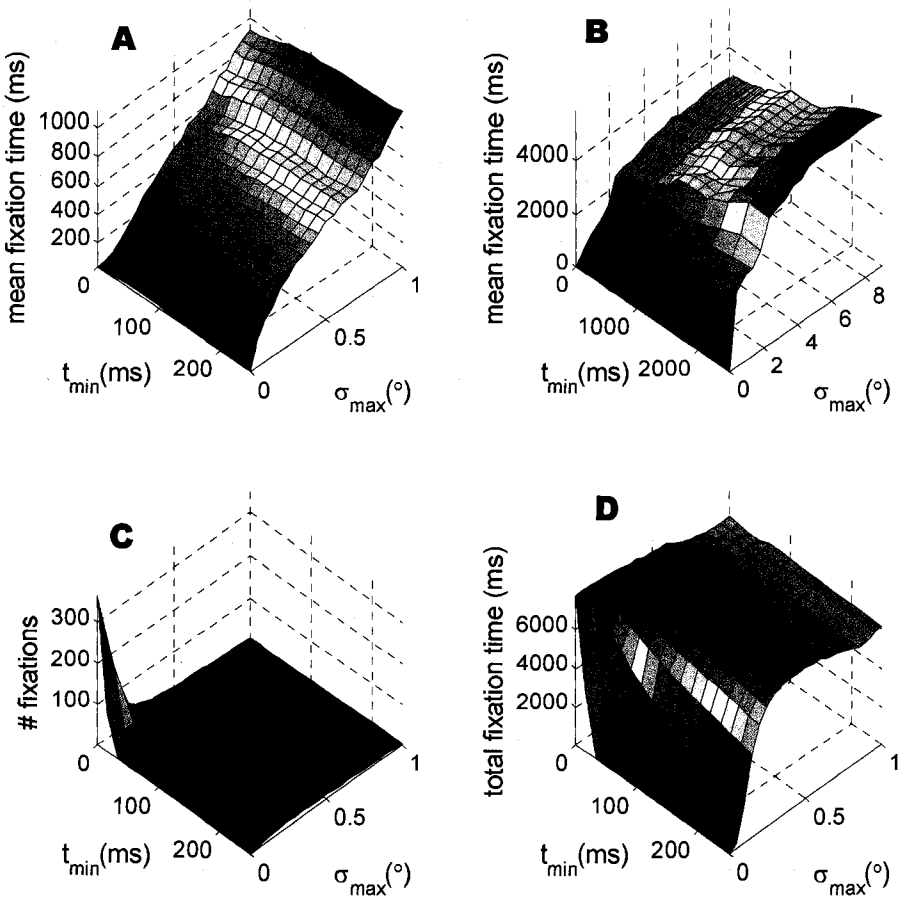


Figure 3.3: Parameter Dependence of Position-Variance Method on Faces (A: mean fixation time; B: mean fixation (extended range); C: # fixations/trial; D: total fixation time/trial). Note that the spatial scale of the position-variance method is approximately 1/6 of the distance algorithm, i.e. 1° position variance ~ 6° distance algorithm.

example, the centroid method's spatial constraint is the distance from the centroid. In comparison, the distance algorithm's constraint is the diameter of the cluster. From this we might assume that the spatial slope of the centroid method should be twice that of the distance method. However, though the centroid method's slope is higher than the distance method, it is not twice as high. This is likely due to the movement of the

centroid as the centroid method expands its fixation window, an effect which reduces the effective scale of the radius constraint in comparison to the distance method. Similarly, the position-variance algorithm employs a measure on the standard deviation of distance from the centroid, a measure which has a natural scale which is much smaller than distance or centroid algorithms. The I-DT algorithm uses a spatial constraint which is the sum of vertical and horizontal spread. This measure is somewhat looser than the distance algorithm, and can cover a larger instantaneous distance. Finally, the velocity algorithm, being the only non-dispersion algorithm, has no natural basis of comparison with the other algorithms. The properties that relate velocity to spatial dispersion are an interesting area of future work.

Note also that the consistent behavior of the algorithms implies that mean fixation duration results for different algorithms can be converted to one another. However, we also note that the relationship between the spatial slopes does not necessarily follow an intuitive pattern and that the scale of each spatial parameter is different, sometimes dramatically (e.g. consider the variance method versus I-DT).

3.2.2 Comparing Scanning on Classes of Objects

The more common use of mean fixation time, however, is for making comparisons. In Figure 3.4 we compare Faces(+) versus Blocks(-) under the distance method. We see that, depending on parameter settings, the mean fixation duration for a particular stimulus class can be either greater or lower than another class. We also see that taking into account the variation in trials also impacts analysis. For example, though the difference

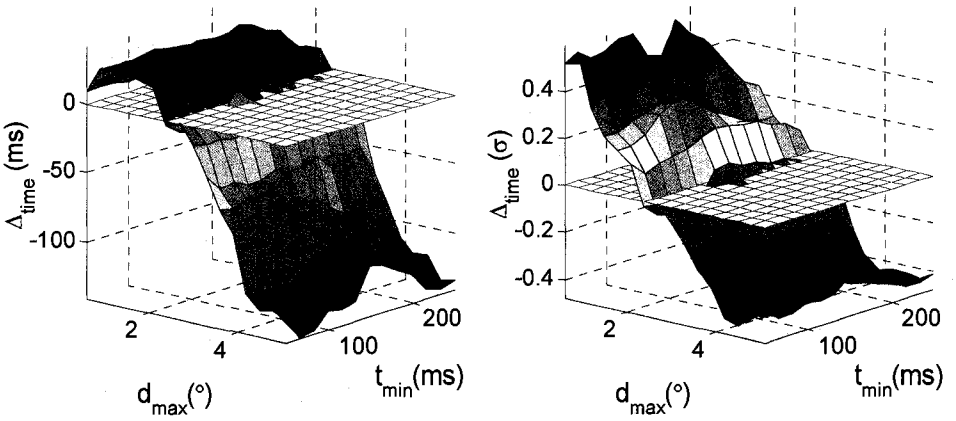


Figure 3.4: *left*: Mean Fixation Time Difference for Distance Method (Faces-Blocks) in ms. *right*: standardized difference defined as the difference divided by the standard deviation. The flat plane in both plots represents the zero surface. When data is above the plane, faces have a greater mean fixation duration; when below, blocks are greater.

in fixation durations at low d_{max} in Figure 3.4 is small in an absolute temporal sense, the effect in comparison to the trial standard deviation is prominent.

In Table 3.2 we examine the regions delineated by the standardized difference (i.e. the difference between the mean fixation times on faces versus blocks divided by the averaged standard deviation). For each method we localize the region corresponding to faces (large positive scores), blocks (large negative scores), and an indeterminate mixed region (small absolute scores). For each of these regions we compute the mean time (t) and spatial (**s**) parameter. For comparability, using the slopes in Table 3.1, we also translate t and **s** to the equivalent distance dispersion algorithm temporal (t_{dist}) and spatial parameters (s_{dist}=d_{max}).

Face Region ($\Delta_{time}(\sigma) \geq 0.2$)

method	t	s	t_{dist}	s_{dist}	Δ_t(ms)	$\Delta_t(\sigma)$	%
distance	133	1.19°	133	1.19°	25.2	.369	18.6
centroid	130	0.77°	145	1.11°	25.1	.350	18.8
variance	136	0.20°	167	1.20°	31.6	.324	17.9
I-DT	143	2.19°	124	1.48°	28.9	.365	24.6
velocity	124	30°/s	73	1.90°	28.1	.349	28.2

Block Region ($\Delta_{time}(\sigma) \leq -0.2$)

method	t	s	t_{dist}	s_{dist}	Δ_t(ms)	$\Delta_t(\sigma)$	%
distance	152	4.71°	152	4.71°	-107	-.339	50.2
centroid	153	2.95°	181	4.28°	-110	-.380	41.5
variance	152	0.64°	165	4.42°	-158	-.520	61.0
I-DT	147	7°/s	112	4.79°	-100	-.326	40.0
velocity	-	-	-	-	-	-	0

Mixed Region ($|\Delta_{time}(\sigma)| \leq 0.2$)

method	t	s	t_{dist}	s_{dist}	Δ_t(ms)	$\Delta_t(\sigma)$	%
distance	156	2.29°	156	2.29°	-1	.016	31.2
centroid	156	1.84°	179	2.65°	-16	-.042	39.7
variance	153	.337°	178	2.17°	-4	-.012	21.0
I-DT	157	4.43°	129	3.00°	-12	-.050	35.4
velocity	160	57°/s	101	3.32°	12	.069	71.8

Table 3.2: Characterization of Regions Corresponding to Differences between Faces and Blocks

Finally, we report the mean temporal differences Δ_t(ms), the mean standardized difference $\Delta_t(\sigma)$, and the percentage of the range space (.6 to 5.1 degrees in the distance algorithm) covered by each region (%).

For all regions, translating spatial parameters to a common scale lead to greater comparability. Differences in mean fixation time were also consistent. The translation of temporal parameters increased variability, however, possibly due to edge effects and the smaller contribution of temporal versus spatial parameters. In terms of coverage, all algorithms were comparable except velocity. This was not unexpected, as dispersion algorithms share common assumptions regarding the spatial cohesion of fixations; the velocity algorithm assumes that fixations are what are left after saccades have been parsed. The two approaches have slightly different operating characteristics which are amplified by differencing.

We found the effects of changing parameters on the final interpretation, as examined by differencing the mean fixation durations of faces and blocks in Figure 3.4 and Table 3.2, to be worrisome. It could be argued that the regime corresponding to higher mean fixation durations for blocks is non-physiological, being centered roughly at the extreme range of foveal vision. However, the results are fairly consistent across all dispersion algorithms, suggesting that rather than being an artifact of data processing, the effects are an inherent property of viewing the respective stimulus classes.

3.2.3 Painting a More Complete Picture with Standard Measures

We thus propose that instead of picking a single set of parameters suitable to all analyses, an approach that is both difficult and incomplete, that the effect of the

parameter space be charted, slopes of effects reported, and different regimes of dominant behavior characterized. This would provide a more complete picture as to the actual scanpath dynamics involved in observing static scenes. Furthermore, given the predictable effect of parameter changes on mean fixation time, this would also allow new results to be compared to the extent literature, and thus could offer a simple method for bridging previously inconsistent results.

Since all the algorithms inherently behave in the same manner, it is an open question as to which algorithm should generally be used in analysis. We note that the ratios of spatial slopes between distance, centroid, and the I-DT methods are preserved across stimulus types. This suggests that this set of algorithms is particularly comparable and that it might be possible to freely convert parameters even without *a priori* knowledge of stimulus changes. Within this set, we believe the distance method is the most transparent and interpretable, as it simply ensures that every point is within some distance to every other point. By contrast, the centroid in the centroid method tends to shift as the fixation is being calculated, and it is unclear as to how this shift impacts the resultant fixation identification. Similarly, the I-DT method is asymmetric with regards to both radial distance and spanned area. For example, a series of points falling within a long, thin region could be viewed as valid as a square of maximal area, despite having an edge nearly twice as long. However, the I-DT method is also the fastest dispersion method examined here and thus could be used when speed is of primary importance.

We should note that the choice of fitting the mean fixation-parameter curves without an offset was done so as to match more closely with the theoretical effects of having a spatial parameter of zero and so as to make the parameter space more

interpretable for the purposes of this study. This leads to some difficulties with the statistics of the regression as it is known that R^2 reported for linear regression without an offset overestimates the fit in comparison to regression with an offset (Gordon, 1981). We address this issue in the next section as we apply these methods to gain some understanding as to the limitations and capabilities of this methodology.

3.3 Applying SLIM

In this section we apply the linear methodology developed in the previous section in order to examine differences between diagnostic categories (children with autism spectrum disorder (ASD), children with developmental delay without autism symptoms (DD), and typically developing children (TD)). Again, we conduct our analysis for several different algorithms (Distance, I-DT, and Velocity), extending our result from the previous section. This section is based on work in (Shic, Chawarska, & Scassellati, 2008b)

3.3.1 Fitting a Plane with an Offset

Participants in this study were 16 typically developing children (TD) (age 25.9±4.7 months), 12 children diagnosed with autism spectrum disorder (ASD) (age 23.9±4.6 months) by expert clinicians, and 5 children diagnosed with developmental disabilities but without autistic syndrome (DD) (age 25.4±5.8 months). All children were matched on chronological age, but ASD and DD children were also matched on verbal mental age (ASD: 14.8±6.5 months; DD: 16.0±8.2 months) as well as non-verbal mental

age (ASD: 20.4±3.3 months ; DD 22.5±7.3 months) as determined by the Mullen Scales of Early Learning (Mullen, 1995). Children were again presented with 6 color images of faces (Lundqvist, Flykt, and Öhman, 1998) and 6 color images of blocks (Figure 3.2). In total, TD children contributed 26 trials on faces and 44 trials on blocks; ASD children 34 on faces and 40 on blocks; DD children 13 on faces and 15 on blocks.

Again, we first ensured that all algorithms and all diagnostic classes were within a comparable regime by pegging them to minimum and maximum mean fixation duration for TD children over the distance algorithm for a candidate set of temporal (50 ms $\leq t_{min}$ $\leq$ 250 ms) and spatial ($0.6° \leq s \leq 5.1°$) parameters. This candidate set served as a reference algorithm. To fit the mean fixation duration t_{fix} we used 3 coefficients: a temporal slope, $slope_t$, a spatial slope, $slope_s$, and an offset, t_0:

$$t_{fix}(t_{min}, s) = slope_t \cdot t_{min} + slope_s \cdot s + t_0 \tag{3.2}$$

The results of this analysis are summarized in Table 3.3.

As we can see, the linear regressions fit the data quite well, with the worst case still accounting for over 90% of sample variance. The good match suggests that converting between algorithms should be fairly straightforward and effective.

In Figure 3.5, we use the coefficients from Table 3.3 to convert all algorithms to a common axis. Note that, in the unscaled graph on the left of Figure 3.5, if the two dispersion algorithms were comparable in terms of parameters, they would directly overlap one another. However, the two surfaces are offset and have different slopes,

Distance-Dispersion Algorithm

	$slope_t$		$slope_s$		t_0		R^2	
diag.	Face	Block	Face	Block	Face	Block	Face	Block
NC	0.67	0.74	142	175	83	-2	.981	.997
DD	0.80	0.78	128	183	86	-15	.922	.987
ASD	0.63	0.57	172	171	21	4	.996	.996

I-DT Algorithm

	$slope_t$		$slope_s$		t_0		R^2	
diag.	Face	Block	Face	Block	Face	Block	Face	Block
NC	0.63	0.64	97	118	64	-26	.989	.997
DD	0.85	0.68	105	116	-6	-13	.934	.985
ASD	0.55	0.49	118	119	-7	-32	.995	.996

Velocity-Threshold Algorithm

	$slope_t$		$slope_s$		t_0		R^2	
diag.	Face	Block	Face	Block	Face	Block	Face	Block
NC	0.95	1.11	11.2	11.4	-138	-205	.984	.983
DD	0.97	1.09	9.6	12.6	-104	-229	.993	.981
ASD	1.04	0.94	13.7	13.5	-244	-271	.914	.949

Table 3.3: Linear regression coefficients ($slope_t$, $slope_s$, and t_0) and regression explained variance (R^2) of mean fixation duration for three different algorithms (distance, I-DT, and velocity), three different diagnostic categories (NC, DD, and ASD), and two different stimulus types (Faces and Blocks).

implying that they are not comparable. In other words, a spatial constraint of 1° for the distance algorithm is not equivalent to a spatial constraint of 1° for the I-DT algorithm.

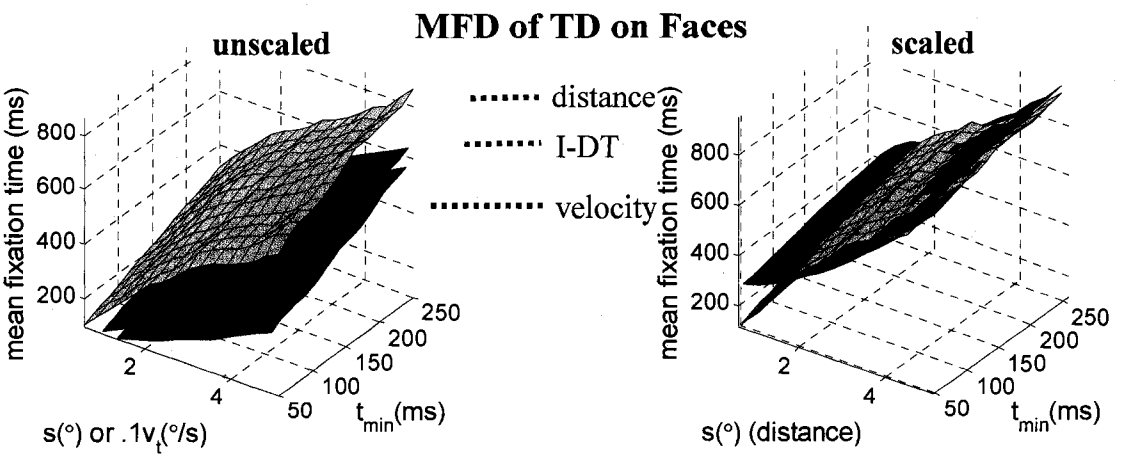


Figure 3.5: Mean fixation duration of TD children viewing faces for different algorithms as a function of spatial and temporal parameters. *Left:* before scaling to the distance algorithm. *Right:* after scaling to the distance algorithm as given by the coefficients in Table 3.3.

Likewise, the velocity algorithm has no natural basis for comparison with other algorithms. Note also that the dispersion algorithms share a somewhat similar scale as they are both in units of spatial degrees, whereas the velocity algorithm has units of degrees per second. For display purposes only, velocity was scaled down by a factor of 10 spatially.

We also can also use our model to simulate versions of mean fixation duration behavior. For instance, in Figure 3.6, top, we examine how TD children differentially treat faces as opposed to the less ecological block designs. In Figure 3.6, bottom, we use a model based solely on the parameters of the regression to generate an idealized version of this behavior. These results suggest that the variation and reversals observed when manipulating fixation identification algorithms is partly due to the natural structure and dependencies of the measure and not some spurious error nor an artifact to be hidden.

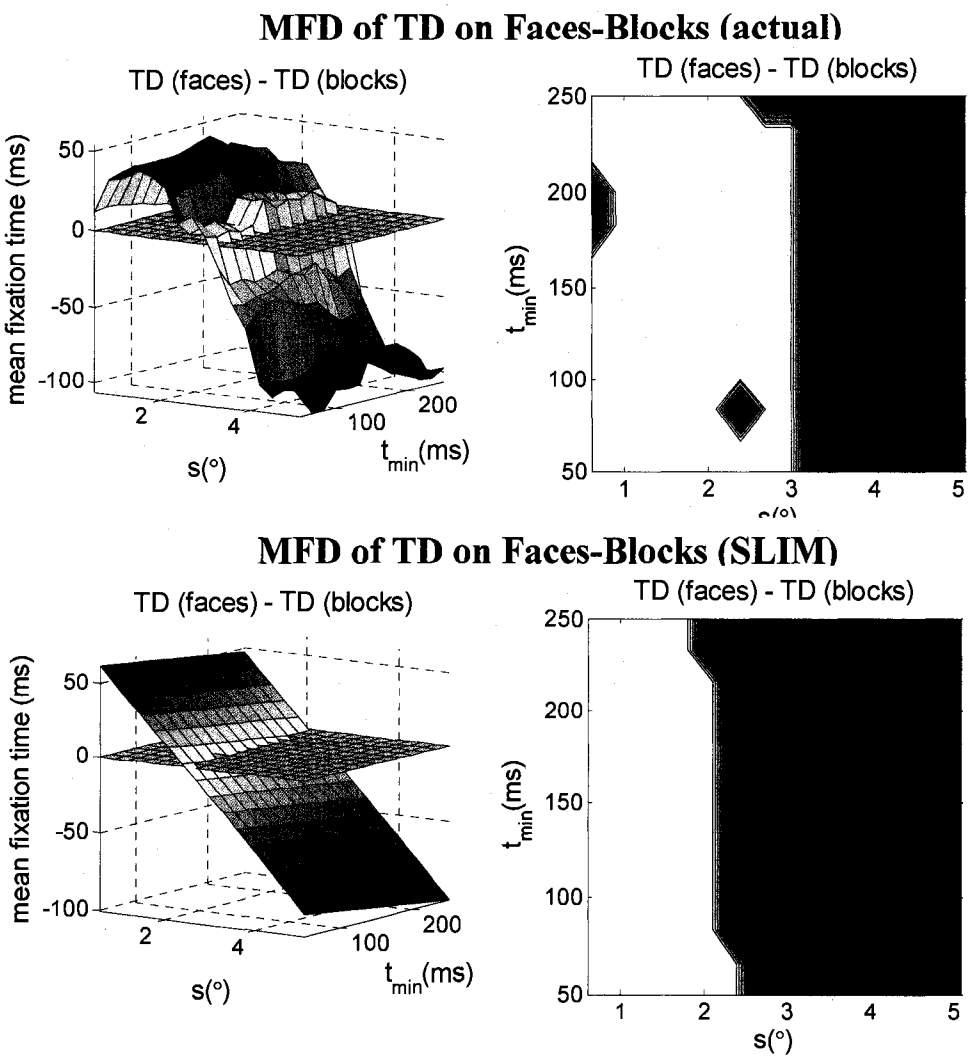


Figure 3.6: *Top:* Differences in mean fixation duration between different stimulus classes (Faces-Blocks) for TD children under the distance-dispersion algorithm. The difference surface is shown on the right. Areas where differences are positive are shown in white on the right, negative in black. *Bottom:* idealized model of the differences built from the regression coefficients in Table 3.3, showing a reversal between theoretical and experimental patterns.

However, the discrepancies between the real and simulated data do also serve as a reminder that the linear model is, in fact, too simple, though it can capture much of the gross, qualitative behavior.

3.3.2 Comparing Scanning across Diagnostic Groups

In order to examine the stability of outcome measures for making comparisons between diagnostic groups, we also examine the difference in mean fixation duration as a function of parameter changes. As there was little effect due to temporal parameter changes, we plot only a representative example at a common t_{min} (Figure 3.7).

We can see that by manipulating the parameters associated with fixation identification algorithms, our reported results can reverse. With one set of parameter choices one group is associated with longer mean fixation durations. With another set of parameters, a different group becomes the group with longer fixations. Notice, however, that the regimes of behavior are fairly large and contiguous, extending to the border of the parameter space. This implies that rather than some random effect, the reversals are tied with some specific spatiotemporal transition.

The reversals of mean fixation duration are quite prominent in the results we have shown. In a traditional analysis, a particular choice of spatial and temporal parameters would be chosen *a priori* and the observed effect would be taken as representative of some global psychological effect. For example, one might look at the low spatial regime of Figure 3.7 while focusing on the fixation duration differences between TD and ASD children (blue line). From just this small slice of the analysis, one might conclude that typical individuals experience a greater cognitive load when observing faces than do

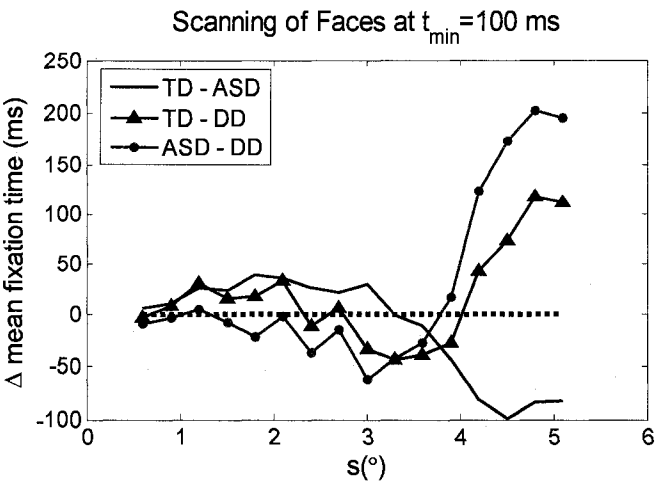


Figure 3.7: Differences in mean fixation duration between diagnostic groups for faces under the distance-dispersion algorithm at t_{min}=100ms. The crossing of the zero-line by comparison lines represent effect reversals.

children with ASD; from this, one might conclude that, for TD children, blocks do not engage more neural mechanisms than faces. However, such analyses miss the larger pattern of behavior which includes the reversals occurring at higher spatial parameter settings. Furthermore, as shown by the essentially flat behavior of algorithms in Figure 3.5, a natural, universal parameter scale for mean fixation duration does not exist. This, combined with the differences observed for stimulus classes, argues against the existence of a unique set of parameters that can be appropriately selected in advance.

The crucial point regarding these parameters is that together the three coefficients (***slope***$_s$, ***slope***$_t$, and t_0) capture the behavior of mean fixation duration quite well. If we examine the behavior of TD and DD individuals in Table 3.3, we find that the there is a modulation of coefficients as the stimulus changes from faces to blocks. This suggests that there is some distributional reaction to the difference in stimuli for these two subject

populations. By comparison, the ASD group is largely invariant to the change. This effect is consistent with known face processing abnormalities and social difficulties in autism, for instance difficulties in unfamiliar face recognition as noted by Boucher & V. Lewis (1992) and Chawarska & Volkmar (2007). Recently, an analysis using the same data set showed that recognition of faces was impaired in the group with autism, but not in typical individuals, and that recognition of blocks was lacking in both individuals with autism and typical development (Chawarska & Shic). This coincides well with the pattern of results we have shown. It is possible that individuals with autism, especially at this young age, view the face in a more pattern-like fashion than their TD or DD peers, unfortunately setting the stage for a cascade of future deficits.

3.3.3 Implications and Limitations

We have shown that choosing a single set of parameters for calculating mean fixation duration is a problematic task, as effect reversals occur both between diagnostic groups and between stimulus classes. We have also shown that no natural comparability exists between different algorithms. However, by computing mean fixation duration over a range of fixation identification parameters, we are able to model mean fixation duration in a straightforward manner. This provides a better understanding of the underpinnings from which differences in mean fixation duration arise, and provides a method for unifying the multitude of disparate fixation identification methods. Finally, the coefficients of our model have given us some way to compare children with autism against controls, showing us that even at the very young ages of the subjects in our study, differences in processing the world are already apparent.

The studies presented in these previous two sections are limited in many ways. First, the populations under study are extremely young children. It is possible that the highly linear effect that we see for mean fixation duration is reflective of the simplicity of early perceptual processing systems. For this reason, this study should be replicated in adults. However, if it turned out that adults did generate nonlinearities that were not found in children, this would be extremely interesting in its own right, as it would imply that some cognitive mechanism coming online was intercepting the more primitive process in children. Furthermore, such a finding would actually strengthen our case for charting the parameter space, because such an effect would likely be poorly characterized by single *a priori* choices in parameter settings. Also, the task that we use is free-viewing embedded within a recognition task. It might be possible that the free-viewing aspects of the experiment are responsible for the simple structure we observe for mean fixation duration and that the imposition of any further experimental structure would break this effect. Again, the limitations under which these effects held would be an interesting avenue of research, and charting the parameters, as we have suggested, could help to pinpoint specific characteristics of certain experimental designs, such as the scale by which long saccades are engaged for the monitoring of multiple well-separated stimuli. In addition, though the subject sample we have chosen is certainly unique, it is small. Notably, there are only five subjects in the DD population. It is our hope that future studies with larger populations and extended experimental conditions will bear out the main results of this study.

Methodologically, there are some omissions. For one, we only consider simple algorithms for fixation identification. We do not consider clustering, optimization, or

more advanced techniques (e.g. see Privitera & Stark, 2000; Salvucci & J. H. Goldberg, 2000; Santella & Decarlo, 2004). However, many of these algorithms also incorporate at least some spatial or temporal free parameters, and therefore their study would also be amenable to the methods presented here. We should note that the data reduction method on which the dispersion algorithms here are based (Widdel, 1984) is greedy and that there may be some benefits to considering all possible fixations and choosing the best match rather than simply the first match. Dynamic programming could make such an overlapping search more efficient and tractable. We also only consider mean fixation duration, though a host of other measures exist (Inhoff & Radach, 1998; Santella & Decarlo, 2004). This was done in the interests of brevity, but also because mean fixation duration is a central statistic in studies examining the relationship between cognitive processing and scene viewing (though for a dissenting view of its utility, see Irwin (2004)).

However, the most glaring omission is that the linear effect that we find to be so ubiquitous across stimuli, diagnoses, and algorithms, is unexplained. Such a simple trend should have some sort of natural underpinning and be derived from either some physiologically constraint or some algorithmic one. In this section we have used the linear trend of mean fixation duration for practical purposes. In the next section, we consider its theoretical basis, one that may be related to power-law scalings in scanning distributions.

3.4 The Fractal Model of Natural Scanning

Many natural phenomena exhibit fractal or self-similar properties. For example, Mandelbrot (1967), in one of the earliest reports regarding the self-similarity of natural phenomena, showed that the length obtained by measuring the coastline of Britain depends on the length of the ruler used to carry out this measurement. For example, in Figure 3.8, we extract the coastline of Great Britain from a postal map ("UK postal areas," 2008) and, over a fixed grid of boxes of particular lengths, count the number of boxes which contain any of the coastline. The number of boxes $N(s)$ of a certain side-length (s) is a power law:

$$N(s) = A\, s^{-\alpha} \tag{3.3}$$

When $N(s)$ is plotted against s on a log-log plot (Figure 3.9), the result is a straight line.

Since Mandelbrot (1967) hundreds of studies have been published demonstrating self-similar or fractal qualities in nature. Fractal properties have been found in the surfaces of sandstone and shale (Wong, Howard, & Lin, 1986), the fracture surfaces of metals (Mandelbrot, Passoja, & Paullay, 1984), in the structure and distribution of rivers and river basins (Rodríguez-Iturbe & Rinaldo, 1997), coasts and continents (Mandelbrot, 1967, 1975), and it seems, everywhere in between, up to the perimeter of interstellar cirrus (Bazell & Desert, 1988) . In ecology, fractal properties have been reported for the flight patterns of albatross (Viswanathan et al., 1996), the foraging patterns of deer and bumblebees (Viswanathan et al., 1999), and even amoebas (Schuster & Levandowsky, 1996). Not content to be confined to the natural world, fractal analysis has been applied to economics (Mandelbrot, 1999), the artwork of Jackson Pollock (Taylor, Micolich, &

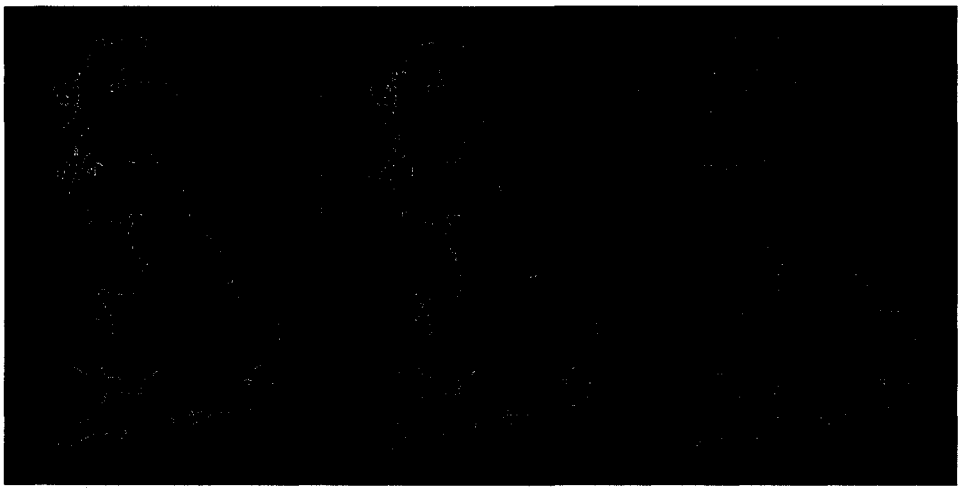


Figure 3.8: Simple Box-counting of the coast of the U.K. The number of boxes needed to cover the coastline increases as a power law of the inverse of the size of the covering boxes.

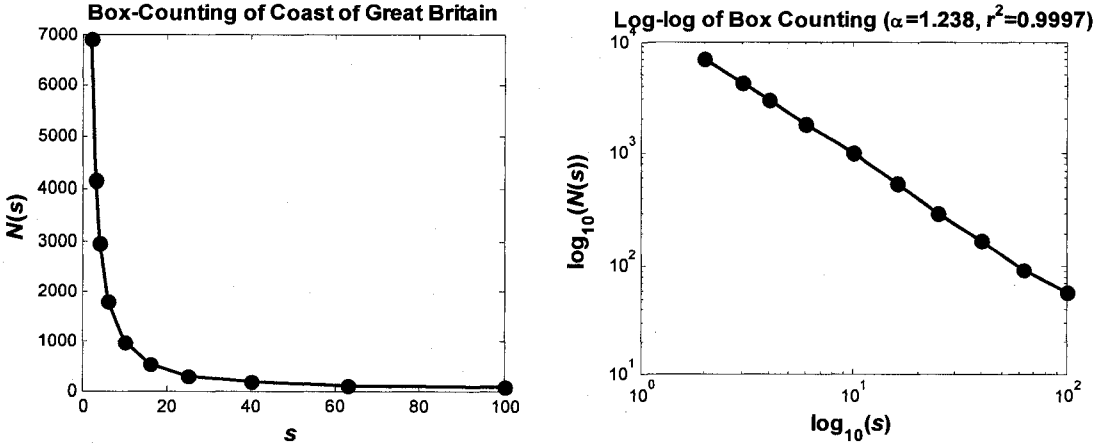


Figure 3.9: *Left*: Number of boxes, $N(s)$, of side-length s necessary to cover the coast of Great Britain. *Right*: log-log plot of left graph, showing a very linear relationship, suggesting a power law for the relationship between $N(s)$ and s.

Jonas, 1999), and investigations into the topology of the internet (Faloutsos, Faloutsos, & Faloutsos, 1999).

The amplitude spectrum of natural images has also been found to obey, on average, a $1/f$ power-law relationship, with f the spatial frequency (Field, 1987; Burton & Moorhead, 1987; Tolhurst, Tadmor, & Chao, 1992) though there is considerable variation across image classes (Torralba & Oliva, 2003). Similarly, the images that we presented to the children in our experiments also have fairly linear, but different, power law relationships in terms of spatial frequency and amplitude (Figure 3.10).

A few research groups have examined power laws in eye-movements. Notably, Brockmann and Geisel have developed a theoretical framework for describing the distributional properties of eye-movements as Lévy flights and show preliminary results that seem to support their model (Brockmann & Geisel, 1999, 2000). Similarly, Boccignone and Ferraro (2004) describe a theoretical model which grounds gaze patterns in terms of low-level features and scene complexity, using a weighted Cauchy-Lévy distribution for jump lengths. Shelhamer, in a series of studies, has demonstrated that predictive saccades exhibit long-term correlations and exhibit fractal properties (Shelhamer, 2005a, 2005b, 2005c; Shelhamer & Joiner, 2003). Aks et al. (2002) report similar findings in a search task when looking at the distance between fixations, a value which is related to saccade amplitude. Liang et al (2005) use detrended fluctuation analysis to examine the scaling exponents of saccade velocity sans microsaccades. It is important to note that all of these studies exclusively focus on the power-law behavior of saccades. As we have shown in previous sections, saccade identification is complicated by problems regarding the choice of parameters. In contrast, we will examine the power law distributions of gaze patterns by characterizing dispersion-based algorithms over multiple scales, obtaining a variant of spatial box-counting suitable for spatiotemporal

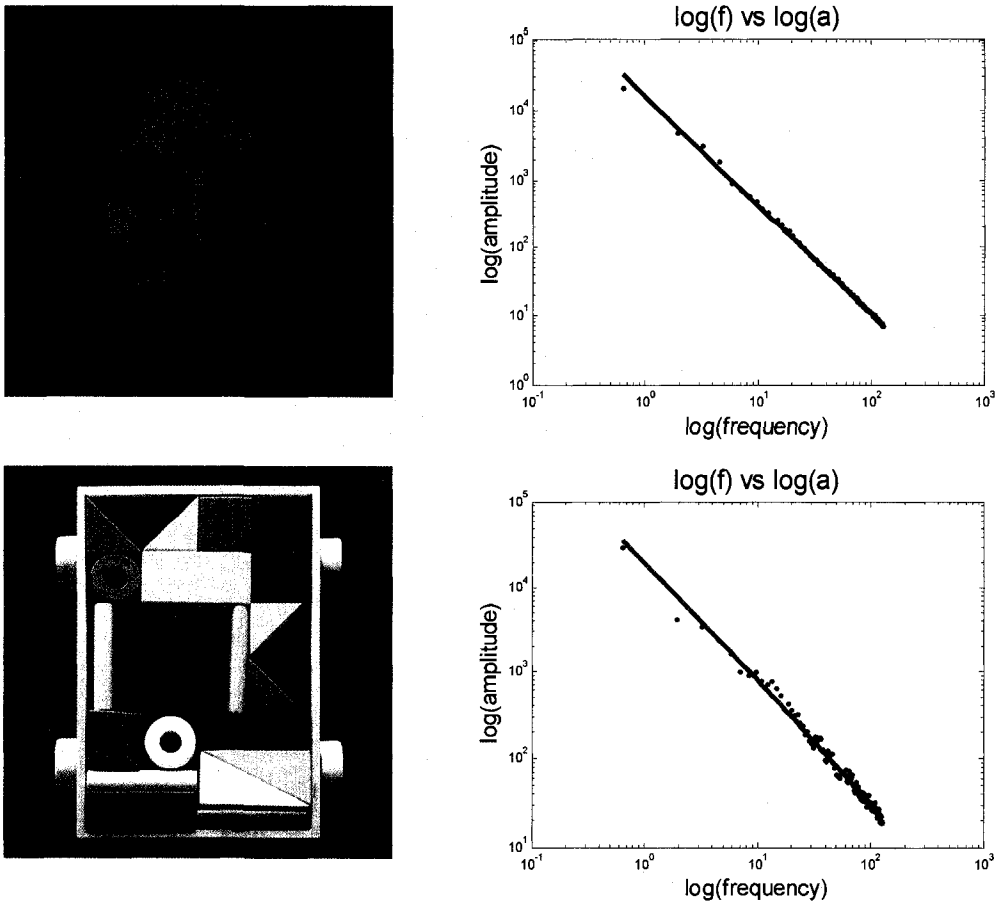


Figure 3.10: Amplitude spectrum of the images used in our study. *Top:* amplitude spectrum for faces, with $\alpha = 1.59$, $R^2 = .998$; *Bottom:* amplitude spectrum for blocks, $\alpha = 1.40$, $R^2 = .987$.

trajectories. We will show that this approach sheds some light on the simple linear interpolation model (SLIM) we have shown in previous sections. This section discusses some material first presented in (Shic, Chawarska, Zucker, & Scassellati, 2008).

3.4.1 Adapting Fixation Measures to Fractal Measures

Though box-counting is typically associated with boxes, the covering object need not be a box (Falconer, 2003; Klinkenberg, 1994). Since it is our intent to examine standard fixation identification algorithms and the implications that the particular distributions of scanning have on these algorithms, we will proceed by examining the operation of the classic greedy dispersion method for fixation identification under the distance spatial constraint. To simplify our analysis, we will neglect the minimum time duration by setting t_{min} = 0ms.

In Figure 3.11 we see how the greedy distance fixation identification algorithm dissects a scanpath. The algorithm begins by identifying a candidate point and then grows as far as it can, temporally, until the next point is d_{max} away from some point already covered by the spatio-temporal cylinder. It then begins at the next valid point and repeats. The total number of cylinders needed to cover the trajectory is the total number of fixations and thus the distance algorithm, viewed from this light, is a temporally-greedy version of box counting.

3.4.2 The Scaling Exponent of Free-scanning in Children

We used the same data from the experiments in Section 3.2 for this analysis. To recap: we presented pictures of faces or block designs (Figure 3.2) to 15 typically-developing children (age 26.5(4.2) months). 4 seconds of valid data were required in the first 10 seconds of stimulus presentation, and only the first 10 seconds of stimulus presentation were analyzed in this study. Subsequently, 46 trials were admitted for blocks and 29 trials for faces (75 trials total). We examined the number of fixations $N(s)$ as a function

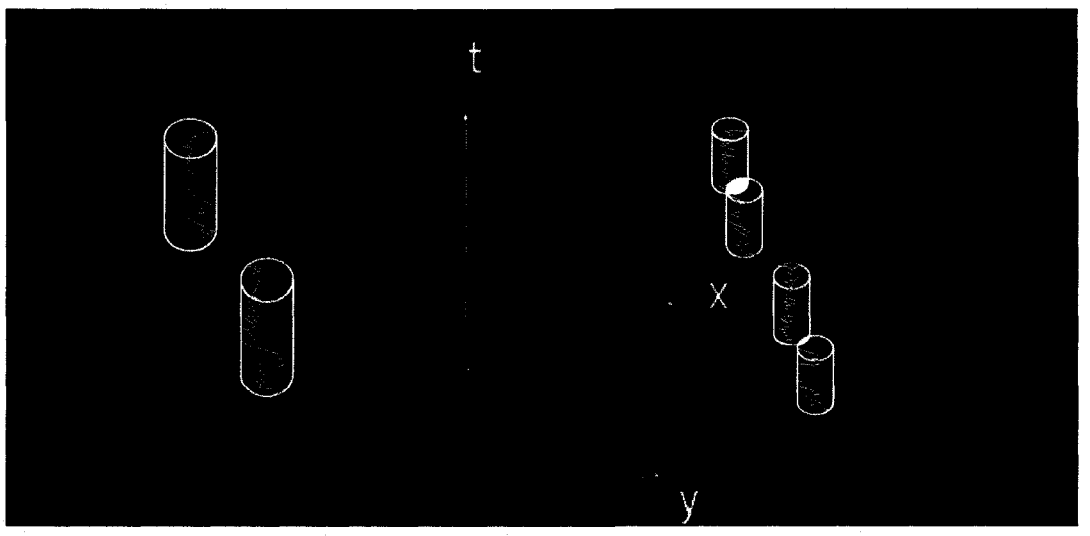


Figure 3.11: Effect of modifying spatial parameter on standard greedy dispersion fixation algorithms. The gaze trajectory is shown moving through time vertically and projects onto the image above and below. *Left*: With a large spatial parameter the cylinder covering the gaze trajectory is thicker, and fewer cylinders are needed to cover the trajectory. *Right*: With a smaller spatial parameter more cylinders are necessary.

of the spatial constraint parameter $s=d_{max}$ under the distance-dispersion fixation identification algorithm. A representative example of one trial is shown in Figure 3.12.

The average R^2 of the regression was .98 (σ=.01), with the minimum fit on any of the 75 trials being R^2 = .94. The scaling exponent α differed (F=4.3, p<.05) between blocks (α =1.28 (.17)) and faces (α =1.19 (.17)). The constant term (A in Equation 3.3) is approximately the log of the total amount of time spent in scanning and also differed (F=7.4, p<.01) between blocks (A = 4.33 (.28)) and faces (A = 4.14 (.33)). A representative comparison is shown in Figure 3.13.

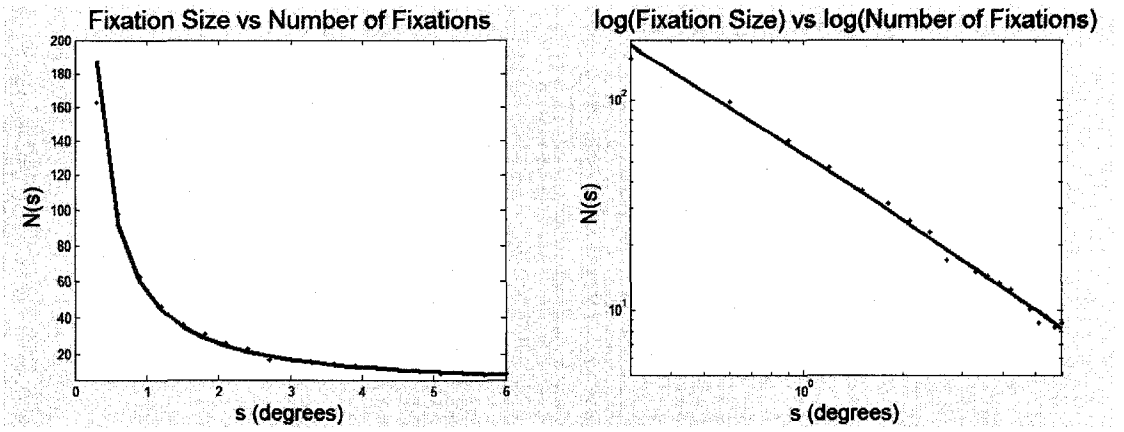


Figure 3.12: Representative example of a single trial. *Left:* Number of fixations $N(s)$ as a function of s, the size of the spatial parameter for the distance algorithm (i.e. the maximal separating distance between spoints). *Right:* log-log plot of the left plot. The line between points shows the theoretical line calculated by least-squares fit of a line to the log-transform of the set of points.

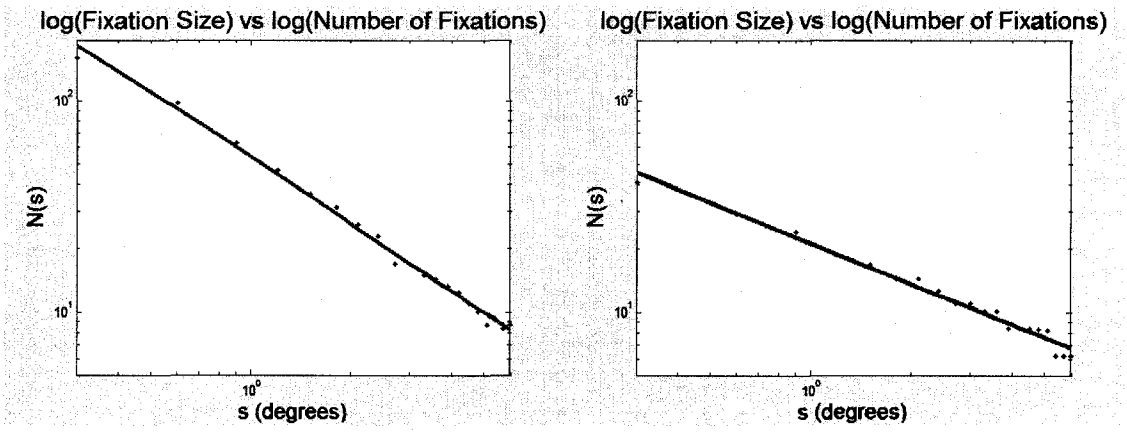


Figure 3.13: Representative example of a single trial. *Left*: log-log plot showing scaling of blocks and *Right:* log-log plot showing scaling of faces. Note the stair-casing on the right plot is due to discretization (i.e. low counts at high spatial scales).

3.4.3 Relationships between $N(s)$ and T_{fix}

If we examine Figure 3.3D, we can see that over a large range the total amount of time spent in fixations appears constant. Given that we have conducted this analysis with no minimum fixation duration for the distance-dispersion algorithm, i.e. $t_{min} = 0$ ms, no data should be lost to saccades when calculating the total fixation duration. Since the mean fixation duration is the total amount spent in fixations divided by the number of fixations, we can approximate the mean fixation duration t_{fix} as:

$$t_{fix}(0,s) = \frac{T_{fix}}{As^{-\alpha}} \tag{3.4}$$

where T_{fix} is the total time spent in fixations, and constant. If this is the case, then the inverse of the number of fixations $N(s)$ should appear linear without log-transformation. This is in fact the case (Figure 3.14).

It appears, then, that the linear trend seen for mean fixation duration in the previous sections is grounded in the distributional aspects of the number of fixations. To further examine how these parameters correspond to distribution aspects of scanning, we conducted a simple simulation examining the model posed in Section 3.3, simple linear interpolation model with offset. We employ an idealized model of saccade generation. In this model, the distribution of saccades is power law distributed:

$$p(a) = ka^{-\beta} \tag{3.5}$$

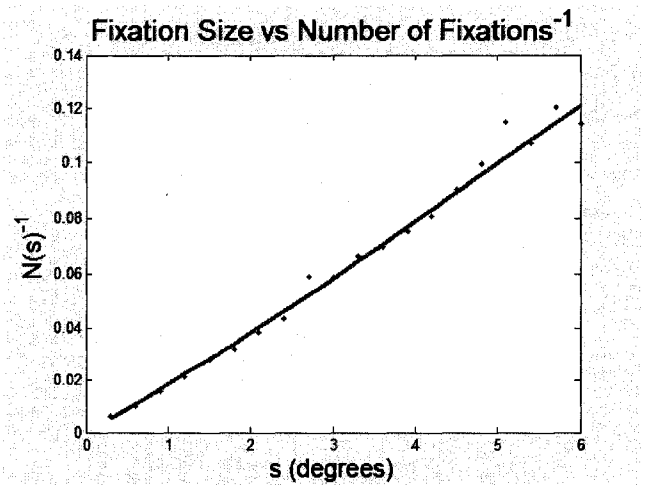


Figure 3.14: Inverse of the number of fixations as a function of the scale of analysis.

where *p(a)* is the probability of a saccade of amplitude a, β is a constant (1.2 to 1.35), and k is a normalizing constant for the discrete range-limited case only. The duration of each saccade is given by the square-root main saccade rule (Lebedev et al, 1996):

$$t(a) = .17\sqrt{a} \tag{3.6}$$

The resulting distributions seem to qualitatively share more similar qualities with real scan patterns than patterns, for instance, generated with normally distributed step-lengths (Figure 3.15).

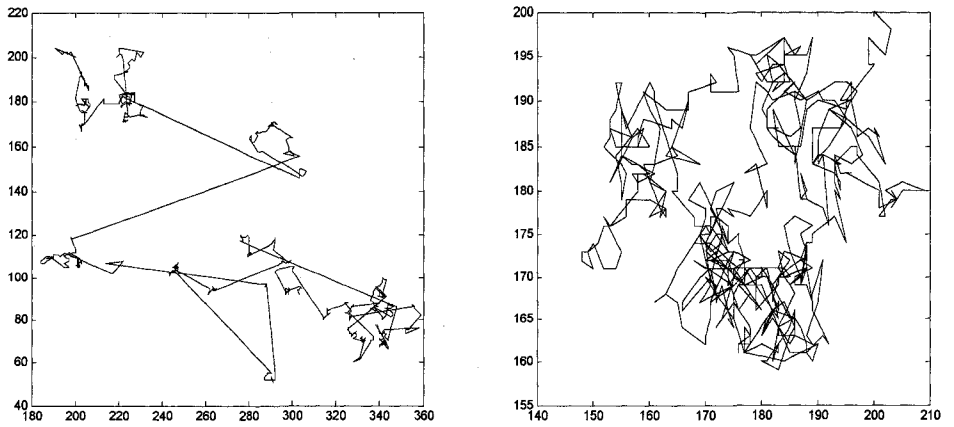


Figure 3.15: *Left:* Generated scanpath with power-law step sizes. *Right:* scanpaths with normally distributed steps.

In terms of numerical correspondences, by increasing α, we increase the spatial slope, $slope_s$ (β =1.2, $slope_s$=132; β =1.4, $slope_s$=179). This suggests that higher spatial slopes correspond to denser scanning patterns.

The temporal coefficient, $slope_t$, characterizes how the mean fixation duration increases as the minimum time requirement t_{min} increases. A larger temporal slope, counter-intuitively, implies a greater loss of data: by removing fixations with shorter durations, the average fixation duration tends to increase. This process explains the discrepancy in temporal slopes for the velocity algorithm as compared to dispersion algorithms (Table 3.3). The temporal constraint for velocity algorithms is a pure rejection criterion; by comparison, dispersion algorithms have a chance to partially recover a fixation as the candidate fixation window slides. In terms of scanpath effects, a larger temporal coefficient implies more non-recoverable short-time fixations, i.e. short time fixations which are separated by large distances. A full analysis of this coefficient

will require an analogous study examining the distribution of the durations of fixations. Should this distribution be found to have a particularly simple form, as we might expect, then the lost time to saccades should be proportional to the integral of the duration distribution up until t_{min}.

The duration offset, t_o, if viewed in the linear domain, could be viewed as a constant added to every fixation regardless of spatial or temporal parameters. However, given the relationship found in terms of number of fixations, it may make more sense to consider a model which does not use a linear offset. We should note that the implications of grounding mean fixation duration with a power-law for the number of fixations would suggest that mean fixation duration is not exactly linear, though for practical purposes it may be well modeled by a plane. In this case, the offsets observed may be associated with errors caused by matching nonlinearities in the mean fixation duration, and should thereby increase as the spatial slope increases. Again, an examination of this effect is a topic for further investigations.

3.5 Grounding the Models

At the current moment it is not possible to fully disentangle the different effects that lead to differences between scanning on blocks and on faces, and the scanning of children with different levels of developmental disability. In this section we will discuss several of the possibilities which may lead to the effects we observe.

The differences in scanning density could be associated with natural image statistics. For instance, Reinagel and Zador (1999) and Parkhurst and Niebur (2003)

show that the local contrast at the point of gaze is higher. Somewhat different results are obtained by Baddeley and Tatler (2006), who show that high frequency edges, but not contrast predict where fixations will occur. In either case, these studies show that scene statistics are biased at the point of regard. It is thus possible that the differences observed in gaze patterns have at least some basis in natural image statistics (for an excellent review of related issues see Geisler (2007)).

Not only are there different image properties in terms of color, contrast, and spatial frequency, faces also recruit higher-order specialized face-specific circuitry (Kanwisher, McDermott, & Chun, 1997). Indeed, there is much evidence to point to the special nature of faces in typical developing infants, children, and in adults (Goren, Sarty, & Wu, 1975; Haan, Pascalis, & Johnson, 2002; Halit, de Haan, & Johnson, 2003; Hoffman & Haxby, 2000; Nelson, 2001; Valenza, Simion, Cassia, & Umilta, 1996). However, we should mention that the privileged role of faces, though found in typical individuals, does not necessary extend to individuals with autism (Hobson, Ouston, & Lee, 1988; Klin et al., 1999; Pelphrey et al., 2002; Pierce, Muller, Ambrose, Allen, & Courchesne, 2001; Schultz et al., 2000). One study found that children with autism performed better, in comparison to a control group, in face processing tasks, when presented with high frequency filtered images of faces rather than low frequency filtered images. Considering that the local spatial frequency of faces differs between locations of the face, it would be interesting to revisit these effects to examine what relationships, if any, could be found with our fractal measures of gaze dimensionality and the statistical distribution of spatial frequency over faces, in these children.

Finally, we should mention that the power-laws that were reported to be found so widely over nature have recently been reexamined by a number of groups. Edwards et al. (2007) examined the power-law behaviors found in the flight or foraging behavior of albatrosses, deer, and bees, and found a number of methodological flaws, concluding that none of these effects were, in fact, power laws. Similarly, a number of groups have criticized what might be construed as an over-exuberance of the power-law phenomena in nature, offering better statistical methods to distinguish between power-laws and other distributions (Bauke, 2007; Clauset, Shalizi, & Newman, 2007; James & Plank, 2007; Shalizi; Sims, Righton, & Pitchford, 2007; White, Enquist, & Green, 2008).

The methods that we have used for power-law fitting in this chapter are quite crude and do not compare with more advanced techniques. In addition, the order of the effect we have examined spans only a twenty-fold scale (corresponding to 1.3 decades or $\log_{10}20$) which does not necessarily correspond with the range which might be expected of a true fractal in nature. However, we will note that this range of analysis coincides with the average found in a survey of work examining fractals in nature (Avnir, Biham, Lidar, & Malcai, 1998). At the same time, though it may be premature to assign a label of "fracticality" to the gaze patterns of children in free-scanning of images, based on methodological reasons, there is evidence to believe power-law distributions are optimal for certain types of foraging and exploration (Viswanathan (1999), though see Plank and James (2008)). It may also be the case that the statistical properties of natural images are well matched to gaze pattern distributions. Given this, it is possible that if power laws are found where hunters hunt for prey, the distribution we see may be explained by the strategy by which the human eye hunts for information.

Even without the designation of fractal, the distribution of the number of fixations as a function of the scale of the analysis does seem to obey a power law. The predictions made by this model go a long way in explaining the effects found for mean fixation duration. Our study suggests that current fixation identification algorithms may be operating under false assumptions about the spatiotemporal distribution of fixations and that there is no single "best" scale for analysis. Our use of fractal dimensionality may offer a route to more robust, more informative, and less biased approaches towards eye-tracking analysis, and leverages common methods that are already in place. It will be interesting to see how other standard eye-tracking measures, once regarded as independent analogues of true physiological events, can be better explained and characterized by the modeling of gaze patterns as distributions. It is through the techniques and approaches demonstrated in this chapter that we hope to finally put the parameter problem in fixation identification to rest.

3.4 Chapter Summary

- We have demonstrated the "parameter problem in fixation identification": that changing the parameters of fixation identification algorithms affects both quantitative and qualitative eye-tracking measures.
- We have shown that the parameter problem is not a problem that can be solved by choosing an "optimal" set of parameters, and that the problem is more a reflection of the distributional characteristics of gaze patterns.

- We have provided a simple linear interpolation model (SLIM) that captures the distributional aspects of gaze patterns for mean fixation duration, and show how the coefficients of this model provide a method for circumventing the parameter problem.
- We have shown that the distributional aspects of the gaze patterns of toddlers with autism, as interpreted through SLIM coefficients, are similar whether these toddlers view faces or non-social abstract block patterns, in contrast to typically developing and developmentally delayed peers.
- We have adopted standard fixation algorithms to perform box-counting, a technique for measuring fractal dimension and have shown that the scanning patterns of typically developing toddlers may have fractal qualities.
- We suggest that scale-free qualities of the scan pattern distribution may be one reason why an optimal set of parameters for fixation identification does not exist.
- We show how characterizing the fractal behavior of gaze patterns can explain why SLIM works as well as it does.

Chapter 4

Region-based Modeling

The standard approach for grounding gaze patterns to specific image properties is through region-of-interest (ROI) based modeling (Duchowski, 2003; Salvucci & Goldberg, 2000). In this approach, the scene under question is carved into a series of usually mutually exclusive regions and measures on the gaze patterns are examined as a function of the regions (Figure 4.1). In this chapter, we present an example of region-based modeling, examining the differences between two year old and four year old children with autism and typically developing controls (beginning in Section 4.1). It is important to note that region based modeling is the de facto standard by which eye-tracking data is analyzed in psychological research. However, general guidelines for how regions should be grouped together for analysis do not exist. We offer a more methodical approach, centered around a strategy of hierarchical analysis (Section 4.2). In addition, the standard approach examines only gross characteristics of looking such as how long individuals spend gazing at each region (Section 4.3). We augment the standard approach by providing a series of dynamic measures for capturing behavior *between* different region (Section 4.4), allowing us to analyze functional circuits of scanning. We will find that these measures are consistent with the standard measures and also provide stronger effects that suggest dynamic analysis may provide greater discriminatory power (Section 4.5). We conclude with a discussion of both the limitations (Section 4.6) and implications (Section 4.7) of our techniques as they pertain to both methodological advances and autism. We note that the exploration in this chapter has been conducted very much in the spirit of the standard

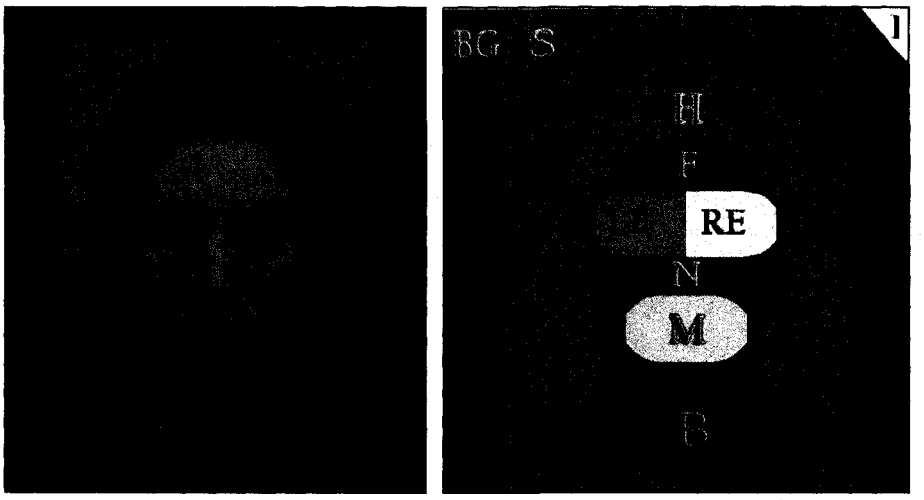


Figure 4.1: Example of regions of interest (ROI) analysis. *Left*: Stimulus image. *Right:* ROIs. The abbreviations for regions are: BG (background), S(background of stimulus), H(hair), F(face), LE(left eye), RE(right eye), N(nose), M(mouth), B(body). I(invalid data)

psychological work in eye-tracking and should be contrasted with our later chapters employing computational methods as a core.

Traditionally, use of ROIs in eye-tracking analysis is not considered modeling. However, the act of cutting up the scene that a subject will see implicitly defines what the experimenter believes is or is not important. In Figure 4.1, for example, we have separated regions of the face as well as regions of the body and background. Though this seems like an obvious way to divide the image, it is important to realize that the designation of these regions is, in fact, somewhat biased by our own preconceived notions of where the natural separations in the images should be. It is not entirely obvious that a newborn infant, or a child with autism, would perceive these separating lines as appropriate. Methods do exist for automatically dividing the scene into regions

algorithmically, based on where patterns of eye-movements occur (Privitera & Stark, 2000), but at the same time this leads to an additional problem in that the regions that are generated need to be interpreted themselves.

We should also note that, in studies where subjects are compliant typical adults, it is rare that the subject will suddenly turn away in the middle of a trial. However, in children this occurs quite often. For this reason, blinks, head turns, and other invalid data should be included into analyses (separated depending on the purposes of the experiments) as we have done in Figure 4.1. In many cases this is as simple as reported the lost time in a trial, or the lost number of trials. Both of these pieces of information can provide valuable clues as to the differences between populations of subjects or performance in tasks (e.g. see Mayes, Bornstein, Chawarska & Granger (1995)).

Furthermore, experimental error can be considerable, and for this reason the regions themselves need to take into account the possible variation due to calibration drift or other technical issues. For example, in Figure 4.1, the ROIs for the eyes have been enlarged to account for possible error. If the region was drawn tight around the eye, for instance, a subject focusing on the corner of the eye might, with the addition of some small error, be considered to be looking away from the eye. Since some regions will naturally impinge on other regions by enlarging, it is important to realize that a not inconsiderable number of design decisions need to be incorporated when selecting regions. It is possible to consider overlapping regions of interest, but this leads to the situation where the total time spent in the parts exceeds the total time spent in the whole, a situation that many researchers might find inelegant.

Even after a scene has been divided into regions, analysis can be quite complicated. It is usually not the case that one wants to consider all the regions at the same time, as this can often lead to a very convoluted and uninformative picture (e.g. consider a bar graph of the 10 regions of Figure 4.1). It is useful in these situations to combine different regions and make comparisons based on a smaller number of clustered areas. For instance, one might consider the left eye and the right eye of Figure 4.1 as simply the eyes, or the set of eyes, nose, mouth, and face skin as simply the face. However, given N original regions, there are 2^N-1 non-empty combinations of those regions. Given that we would like to consider more than one region at a time, the number of possible ways for assembling the analysis becomes astronomical. This is an argument very similar to the one made by Tsotsos (1988), where he shows that to consider all the relationships necessary for perception in the straightforward manner would require an astronomically large brain. His solution for this situation is a processing architecture that incorporates 1) hierarchical organization; 2) scene topography optimizations; and 3) pooled response, which implies certain features can be aggregated rather than communicated directly. His computational model is a model of visual attention via selective tuning (Tsotsos et al., 1995). What is needed for analysis is a model for selective analysis.

We propose that a good methodology for the region-based analysis of eye-tracking data follows the points of Tsotsos. Namely, we will 1) process the regions in a top-down hierarchy, starting from the coarsest possible view; 2) combine regions only when they are topographically adjacent or semantically similar; 3) use summary measures rather than individual constituents. The following sections are based on

research first presented in (Shic, Chawarska, Bradshaw, & Scassellati, 2008) and illustrate the principles we have discussed.

4.1 ROI Analysis: Face Processing in Toddlers with ASD

51 children participated in this study, divided between two diagnostic categories and two age groups. The younger age group consisted of 12 toddlers with autism spectrum disorder (ASD) (mean age 25.6(5.6) months) and 13 typically developing (TD) toddlers (mean age 25.1(6.0) months). The older age group consisted of 13 children with ASD (mean age 43.6(5.4) months) and 13 TD children (mean age 45.0(4.3) months). Diagnosis of ASD was obtained at the time of testing through standardized assessment instruments (Lord, 2002; Mullen, 1995) and expert clinical observations. The children in this study were on average fairly impaired, having a non-verbal mental age developmental quotient of 80 and a verbal mental age developmental quotient of 61 as determined by the Mullen Scales of Early Learning (Mullen, 1995). For further details and a discussion regarding the stability of early diagnosis, see (Chawarska, Klin, Paul, & Volkmar, 2007).

Children were presented with 6 color images of faces (Lundqvist et al., 1998) (Figure 4.1) centered at a distance of 75 cm from the centerline of the children's eyes. Experiments for Age Group 1 (the younger group) were conducted on a 20" widescreen LCD monitor, such that the stimulus (including grey background) measured 12.8° wide by 17.6° tall. Experiments for Age Group 2 (the older group) were conducted on a 24"

widescreen monitor, such that that the stimulus measured 15° wide by 21° tall. We will later return to these differences.

Stimuli were presented until the child had examined the image for a total of 10 seconds (as determined on-line by a trained experimenter) as part of a Visual Paired Comparison protocol (VPC) (Fantz, 1964). This procedure proceeds by first showing a subject an image to be familiarized. Then, once familiarization has been completed, a recognition phase begins where the subject is shown images side by side, one being the familiarized image and another being a novel image. Preference away from chance indicates that some aspect of the scene has been encoded. Familiarization typically lasts until the child has looked at the scene for a preset amount of time, not counting the time that the child is looking away. However, in this study we only examined the first 10 seconds of eye-tracking data (whether or not the child was fully attentive) in order to collect comparable information regarding inattention to stimuli. We also did not examine the recognition phase of the VPC.

Quality measures included checks on the consistency of eye-tracking calibration and a requirement that the 10 seconds of looking be acquired before a cutoff of 20 seconds. Only children who passed quality tests for at least half of the stimuli presentations were retained in this study. The 51 subjects of this study represent those children who met all data quality criteria (approximately 80% of the initial pool of subjects). Results in this study are presented at the subject level, with each subject's measures averaged over valid trails.

4.2 Hierarchical Analysis

In order to manage the combinatorial explosion of comparisons we conducted a progressive regional analysis of the children's scanning patterns. This progressive analysis started at the highest level, including all possible regions, and gradually zoomed in on more information-dense facial regions. Three levels of analysis were employed. The top level (*Level 3*) began with an examination of the gross characteristics of attention, comparing scanning away from the main stimulus (*non-stimulus regions*: invalid data, stimulus background, screen background) with attention towards the stimulus (*stimulus regions*: eyes, mouth, skin areas, nose, hair, body). This analysis was followed by a mid-level of analysis which tapped information extraction from faces (*Level 2*), comparing the scanning of information-poor regions of the face (*non-key regions*: skin areas, nose, hair, body) with information-rich features (*key regions*: eyes, mouth). The final, and lowest, level of analysis focused on the canonical face processing circuit (*Level 1*) between the eyes (*eye regions*: left eye, right eye) and the *mouth* region.

4.3 Static Time Analysis

Measures for static analysis included the time spent in each region. Note that by "static" we do not mean to imply that the behavior of the scanning patterns, or their distributions, are time invariant, but rather that the statistics used do not carry information regarding transitions or dynamic behavior. Also note that there are other measures that could be included here, notably the ratios of times spent in different regions (Chawarska & Shic).

4.4. Dynamic Time Analysis and Entropy Measures

Measures for dynamic analysis included the number of transitions between *outer* (less informative) and *inner* (more informative) regions. We also considered the entropy H (in base 2, i.e. in "bits" of information) of transition ratios in the three-stage functional circuit spanned by the *outer* area and the two *subregions* of the *inner* region (*Level 3: non-stimulus, non-key, key; Level 2: non-key, eyes, mouth; Level 1: mouth, left eye, right eye*):

$$H(R) = -\sum_{r_i \in R} p(r_i) \log_2 p(r_i) \tag{4.1}$$

where R is set of transitions under consideration and $p(r_i)$ is the ratio (probability) of taking a particular transition r_i belonging to R (Cover & Thomas, 2006). We did not examine the two-stage *outer-inner* circuit (e.g. *non-stimulus* vs *stimulus* in *Level 3*) because in this case entropy provides the same information as a ratio. Typically, entropy is associated with randomness. However, in this context, entropy reflects a more even distribution of transitions between different regions. It is thus more closely aligned with a preference for exploration.

We also conducted a Markov chain entropy analysis of the scanning patterns of the three-stage *outer-inner-subregions* circuit (e.g. *non-stimulus, key,* and *non-key* in *Level 3*). For each trial, we computed an approximate Markov matrix for that trial through sampling. We then characterized the entropy of the matrix. Given a kth order Markov matrix M with transition probabilities m_{Xi} where

$$m_{Xi} = p(X_{n+1} = r_i \mid X) \tag{4.2}$$

and where X is a kth order history of past states (Cover & Thomas, 2006), e.g. $X = \{X_n = eyes, X_{n-1} = mouth, \ldots, X_{k-1} = mouth\}$ we can compute the entropy $H(X)$ as:

$$H(X) = -\sum_{r_i \in R} m_{Xi} \log_2 m_{Xi} \tag{4.3}$$

and the entropy of matrix M, $H(M)$, as:

$$H(M) = \sum_{X \in \mathbf{X}} p(X) H(X) \tag{4.4}$$

with $\boldsymbol{X}$ the set of all possible histories. We do not model self transitions because they dominate in a frame by frame analysis, confounding switching rates with timing. Technically, the model we consider is a semi-Markov chain ignoring dwell. For brevity, we refer to these semi-Markov models as Markov. Similarly, the entropy above is the conditional entropy (dependent on history), which we simply refer to as entropy.

There have been several methods that have used variations on entropy as a measure in eye-tracking. Kruizinga et al. (2006) employ an entropy measure on a single row of the transition matrix to calculate the entropy of specific ROIs. Boccignone and Ferraro (2004) use a 1st order entropy method to calculate the complexity of total scanning over a grid of regions. Althoff and Cohen (1999) combine the entropy of

matrix cells with row and column entropy totals, normalized by the column entropy total, to obtain a measure they term S1 and S2. Our method proceeds similarly and is quite simple, as it simply reflects the entropy rate of a discrete Markov chain (Cover & Thomas, 2006). We will see that this measure is sufficient to generate natural results that are quite interpretable. As it is derived from fundamental information theory, there is also a rich basis for expanding its application.

Note that it is not possible to move between non-adjacent regions without crossing interim areas. For example, to scan between eyes and mouth, one must pass through the face. We have considered both raw unadulterated streams as well as streams with transitions lasting 50 ms or less (transients) removed. Transition counts and ratio entropies are transient-removed since they are easier to interpret when saccade effects are mitigated. This process did not alter our general pattern of results. For Markov matrices, we used raw input streams since transients can be accommodated by increasing model order. States not under consideration were removed from analysis.

The information contained by transition rates, entropy of transition ratios, and entropy of the Markov matrix are complementary. Transition rates give an overview of how often movement is occurring between regions. The entropy of transition ratios marks how skewed the distribution of transitions are. The Markov matrix is a fine level frame-to-frame predictive model of scanning. Note that the transition ratio entropy does not account for directional asymmetries whereas the Markov entropy does. For example, a clockwise pattern *left-eye*, *right-eye*, *mouth* back to *left-eye* might have a high transition ratio entropy, since transitions occur at equal frequency, but a 1st order Markov entropy would show that the pattern is essentially deterministic (zero entropy).

4. 5 Results of Hierarchical Analysis

We have summarized the results of static analyses in Table 4.1 (time), the results of dynamic transition analyses in Tables 4.2 (counts) and 4.3 (entropy), and the results of Markov model analyses in Table 4.4. The reported results are based on multiple analyses of variance (MANOVA) with between-subject factors age and diagnosis.

Region	age group 1		age group 2	
	ASD	TD	ASD	TD
NonStim	3828 (600)	4612 (666)	5539 (535)	2429 (220)
Stim	6172 (600)	5388 (666)	4461 (535)	7571 (220)
NonKey	2150 (264)	1803 (269)	1628 (233)	2575 (240)
Key	4022 (466)	3586 (473)	2834 (406)	4995 (294)
Mouth	618 (150)	1151 (328)	535 (110)	1589 (303)
Eyes	3404 (483)	2434 (539)	2299 (348)	3406 (315)

Table 4.1: Static Analysis - Time Spent in Region (ms).

Transition Regions	age group 1		age group 2	
	ASD	TD	ASD	TD
NonStim – Stim	5.14 (.54)	5.60 (.62)	6.84 (.64)	5.81 (.93)
NonStim – NonKey	2.22 (.36)	2.08 (.40)	3.11 (.32)	2.34 (.39)
NonStim – Key	2.92 (.35)	3.52 (.46)	3.73 (.50)	3.47 (.57)
NonKey – Key	4.36 (.59)	3.80 (.56)	3.20 (.47)	5.79 (.52)
NonKey – Eyes	3.40 (.50)	2.43 (.52)	2.42 (.46)	4.05 (.42)
NonKey – Mouth	.96 (.29)	1.37 (.43)	.78 (.21)	1.74 (.37)
Eyes – Mouth	.88 (.18)	.76 (.17)	.89 (.31)	1.91 (.35)
L.Eye – R.Eye	1.46 (.38)	1.46 (.43)	1.28 (.38)	1.55 (.32)

Table 4.2: Dynamic Analysis - Number of Transitions (count)

Transition	age group 1		age group 2	
	ASD	TD	ASD	TD
Level 3	1.39 (.05)	1.42 (.04)	1.50 (.02)	1.39 (.05)
Level 2	1.04 (.16)	1.06 (.12)	1.06 (.13)	1.29 (.08)
Level 1	.88 (.14)	.76 (.16)	.68 (.19)	1.22 (.07)

Table 4.3: Dynamic Analysis - Entropy of 3-stage Level Circuit (bits)

Level	Order	age group 1		age group 2	
		ASD	TD	ASD	TD
3	*0*	1.480 (.013)	1.437 (.032)	1.522 (.010)	1.428 (.025)
3	*1*	0.721 (.032)	0.651 (.041)	0.810 (.022)	0.607 (.052)
3	*2*	0.578 (.034)	0.546 (.041)	0.678 (.028)	0.523 (.043)
2	*0*	1.314 (.029)	1.246 (.052)	1.292 (.028)	1.367 (.022)
2	*1*	0.414 (.041)	0.374 (.050)	0.424 (.030)	0.519 (.031)
2	*2*	0.247 (.041)	0.216 (.036)	0.244 (.030)	0.367 (.032)
1	*0*	1.272 (.041)	1.185 (.074)	1.265 (.065)	1.391 (.025)
1	*1*	0.166 (.043)	0.094 (.029)	0.167 (.029)	0.257 (.030)
1	*2*	0.064 (.017)	0.026 (.012)	0.045 (.015)	0.116 (.018)

Table 4.4: Dynamic Analysis - Markov Chain Entropy (bits)

4.5.1 Level 3 (Top Level): Attention and Motivation

Here we consider the functional circuit consisting of *non-stimulus* regions *key* and *non-key stimulus* features. A 2 (age) x 2 (diagnosis) analysis of looking time at the *non-stimulus* area indicated a significant effect of diagnosis ($F(1,50)=4.8$, $p<.05$) and an age x diagnosis interaction, ($F=13.4$, $p<.001$) (see Figure 4.2a). Older TD children looked more at the *stimulus* than the younger group, ($F=9.7$, $p<.01$), but in ASD the pattern was

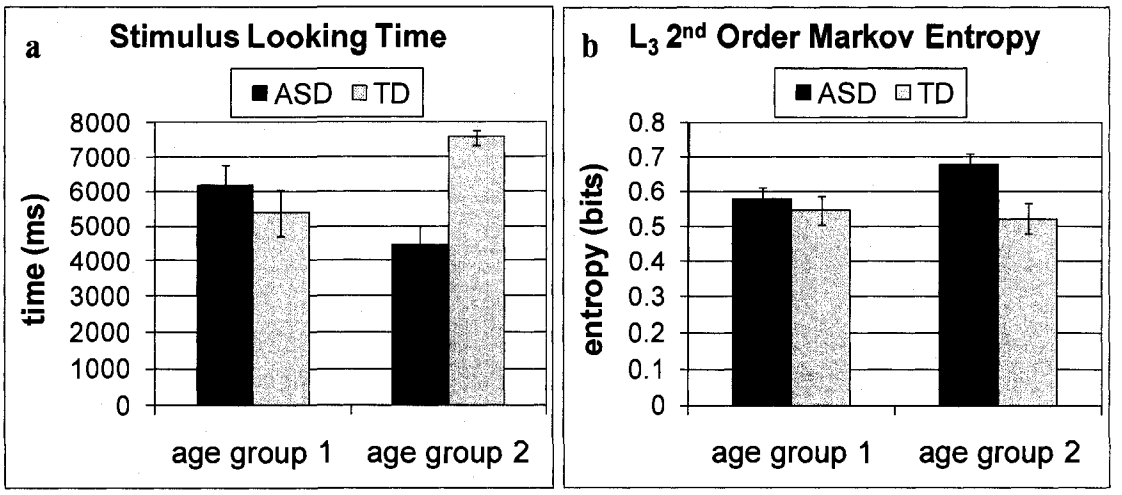


Figure 4.2: Measures from Level 3 circuit: (a) total average *stimulus* looking time (b) Markov matrix entropy of *stimulus - non-key - key* circuit. Error bars are standard error.

reversed, with older children looking less at the *stimulus* than younger ones ($F=4.6$, $p<.05$).

There were no significant effects of age and diagnosis on the number of transitions between *non-stimulus* and *stimulus*, which was surprising, given the discrepancy noted for times (see Table 4.2). However, there was an effect for diagnosis on the entropy of the Markov matrix for all orders (0^{th}: $F=9.7, p<0.01$; 1^{st}: $F=12.3$, $p<0.001$; and 2^{nd}: $F=6.3$, $p<0.01$) (see Figure 4.2b for the characteristic effect for 2^{nd} order Markov, Table 4.4 for others), with the entropy of TD children universally lower than the entropy of ASD children. This suggests that there is less "exploration" of the *non-stimulus – stimulus* circuit by TD children, i.e. ASD children are making proportionally more transitions to and from *non-stimulus* states. It is important to note the difference between transition count and entropy calculations: the former is a raw

tabulation which can be highly variable depending on the child; the latter is a relative measure which takes the context of the scanning circuit into account.

4.5.2 Level 2 (Mid-Level): Face Saliency

Here we consider the circuit spanning *non-key* and the *key* features of the *eyes* and *mouth*.

Analysis of looking times at the *key* features indicated a significant effect of diagnosis ($F=4.3$, $p<.05$) (Figure 4.3a), and a significant age x diagnosis interaction, ($F=9.8$, $p<.01$). For the *non-key* area, there was only a significant age x diagnosis interaction, ($F=6.6$, $p<.05$). A planned comparison revealed that this effect was driven partially by a significant increase in looking at both regions for TD (*non-key* features: $F=4.3$, $p<.05$; *key* features: $F=.6.4$, $p<.05$). These results taken together suggest that the amount devoted to *key* and *non-key* regions at age group 1 by both ASD and TD children is quite similar; similarly, ASD behaviors do not change for these regions between the two time points. However, there is a significant increase in looking at the face, and, in particular, at critical areas of the face, in TD children.

We also compared the number of transitions between *non-key* areas and the *mouth* and *eyes* areas (Figure 4.3b). TD children transitioned between *non-key* areas and the *mouth* significantly more frequently than ASD children ($F=4.1$, $p<.05$) (see Table 4.2). Analysis of the number of transitions between *non-key* areas and the *eyes* revealed only a significant age x diagnosis interaction ($F=9.7$,$p<.01$) (Figure 4.3c), suggesting an increase in frequency of shifts between *non-key* and *eye* areas in older TD children ($F=7.3$, $p<.05$), but not in the ASD children.

For our Markov transitions, there was a main effect of both age and age x diagnosis on entropy for order two models (respectively $F=4.5$, $p<.05$; $F=4.8$, $p<.05$) with TD children having a greater entropy in age group 2 as compared to age group 1 ($F=9.7$, $p<.01$) (Figure 4.3d) and no change in ASD. Again, this is consistent with the notion that TD children explore critical areas in a less deterministic fashion as they get older. By contrast, exploration measures in ASD do not differ between two years and four years of age.

4.5.3 Level 1 (Ground Level): Canonical Scanning

Here we consider the circuit spanning the *mouth* and the *left eye* and the *right eye.* A 2 (age) x 2 (diagnosis) analysis of looking time at the *mouth* area indicated a main effect of diagnosis ($F=10.5$, $p<.01$), with TD children looking at the *mouth* more than their ASD peers (Fig. 4a). An analogous analysis on the *eye* region indicated a significant age x diagnosis interaction ($F=5.8$, $p<.05$). This interaction was due to TD children, in age group 2, looking at the eyes for longer periods than at age group 1 (Fig. 4b). There was no significant difference in looking time at the eyes between the two ASD groups.

For transitions, there was a significant effect of age on the number of transitions between the *eyes* and the *mouth* ($F=4.9$, $p<.05$) as interaction ($F=6.0$, $p<.05$) (Fig. 4c). Again, there was a significant increase in the number of transitions from age group 1 to age group 2 in TD children ($F=9.9$, $p<.01$), but not in the ASD groups. Additionally, there was an interaction of age x diagnosis on transition entropy for the full circuit of *mouth*-(*left-eye*)-(*right-eye*) ($F=7.3$, $p<.01$) caused by an increase in entropy for TD children at age group 2 as compared to age group 1.

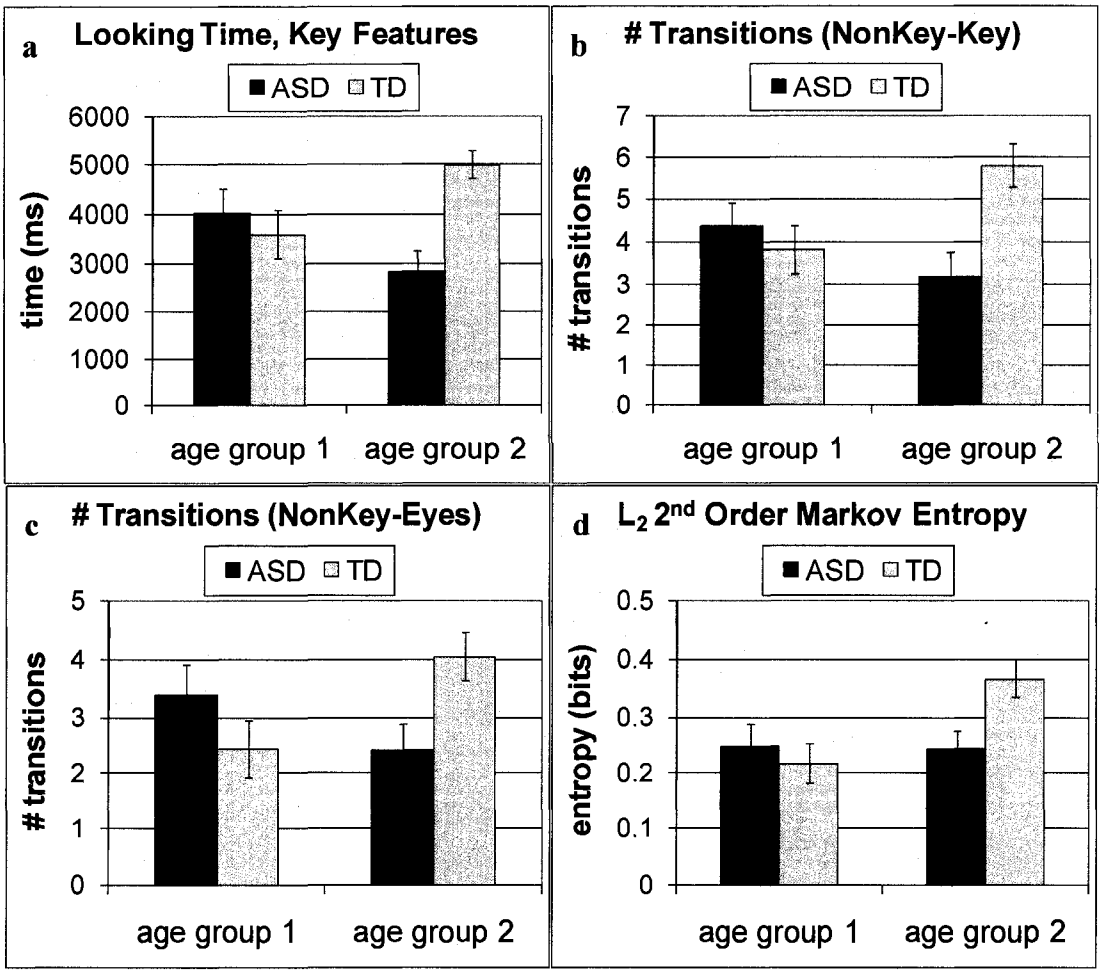


Figure 4.3: Measures from Level 2 circuit: (a) total average looking time at *key* features (b) average number of *non-key – key* transitions per trial (c) average number of *non-key – eyes* transitions (d) Markov matrix entropy of *key - mouth - eyes* circuit.

For Markov analysis (Figure 4.4d), we found a significant effect of age on entropy for both the 1st order and 2nd order chain ($F=6.27$, $p<.05$; $F=5.3$, $p<.05$, respectively) as well as significant interaction effects ($F=6.1$, $p<.05$; $F=12.3$, $p<.001$). There were no differences between older and younger ASD children for either 1st or 2nd order chains. However, there was a significant increase in entropy for both 1st order and 2nd order chains in TD controls ($F=15.6$, $p<.001$; $F=17.3$, $p<.001$).

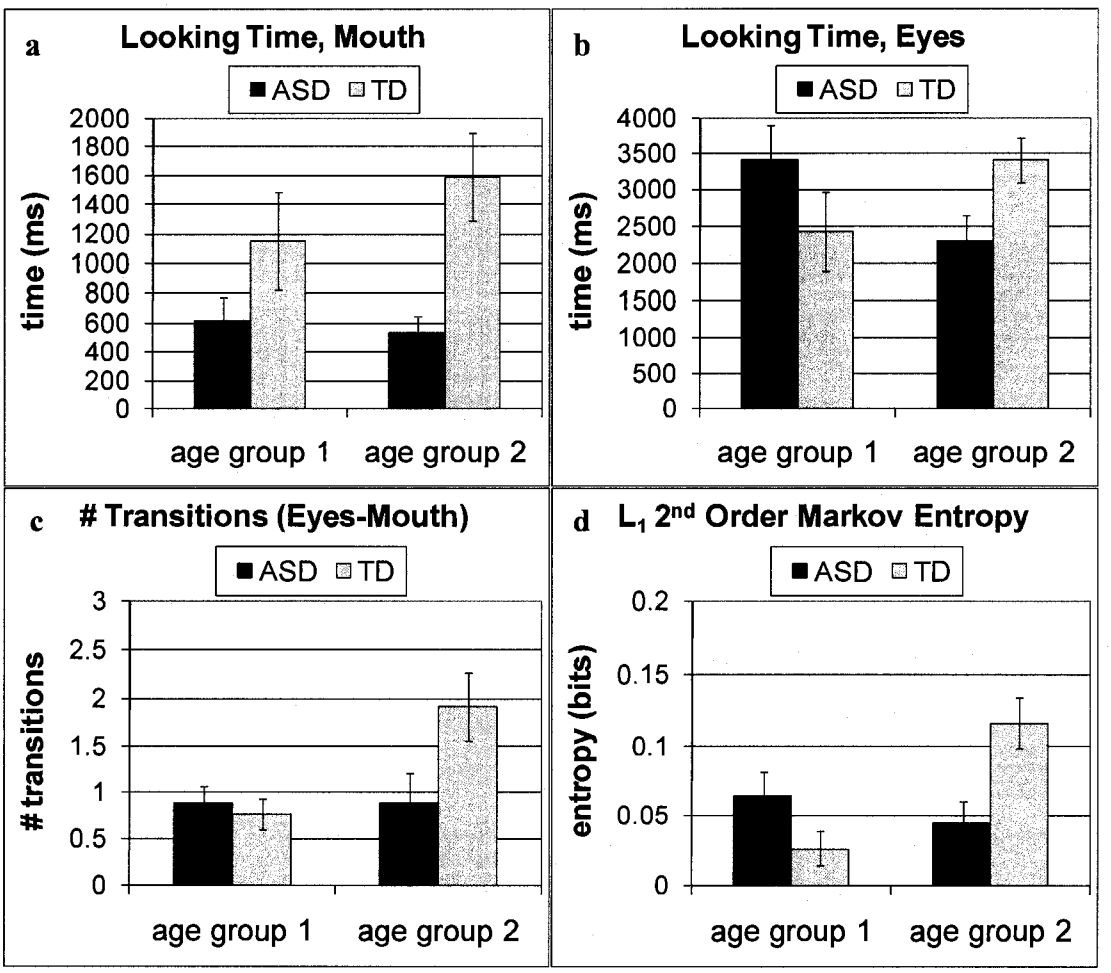


Figure 4.4: Measures from Level 1 circuit: (a) total average *mouth* looking time (b) *eyes* looking time (c) number of *eyes-mouth* transitions (b) Markov matrix entropy of *mouth - left eye - right eye* circuit.

These results, taken together, suggest a functional aberration from typical development in canonical facial feature scanning in children with ASD. At a time when typical children seem to be gravitating towards focusing on core features such as the mouth and eyes, children with ASD are found to be unchanging in their strategies from age group 1 to age group 2.

4.6 Limitations

Several limitations of this study need to be noted. First, the design of our trials is cross-sectional and thus we cannot be sure that the age groups are completely comparable. Selection criteria for valid trials and choices for data stream analysis were made on the basis of methodological concerns. Furthermore, we have not measured any aspects of drop-out.

We also did not include a mental-age matched sample for our ASD population (e.g. a developmentally delayed group). Thus, the results we have found could be attributed to differences in cognitive functioning. Similarly, it is not clear whether the lower attention towards faces found in older children with ASD reflects increasing social deficits or increasing attentional difficulties in general.

The use of two different screen sizes for the older and younger populations is also a potential confound. However, within age group comparisons do not suffer from this deficit and TD and ASD individuals typically do not behave in the same direction (i.e. both up or both down at time 2 from time 1). Our analysis has also indicated no significant effect of the monitor size on scanning patterns.

It is also important to note that many trends apparent in the graphs, such as decreased looking times at *eyes* in older children with ASD, were not significant. It is possible that with a larger sample size these trends could become more prominent, changing the interpretation of our results.

Finally, our measures are fairly new and we have only begun to examine their limitations, interpretations, and interrelationships. It is critical to note that, though the theoretical basis by which entropy measures are employed in signal processing and

communication applications are well known and have been long studied, a formal step-by-step investigation into their use in eye-tracking has not yet been completed. It is likely that the measures employed in this study would benefit greatly by much more controlled experiments relating psychological and cognitive phenomena to the measures we have employed. Additionally, we note that in the limit of very rare transitions, such as is found in the eyes, the entropy calculations used for eye could be significantly improved by aggregating over the entire population rather than at the level of a single subject. Future work will consider technical issues such as the effect of data loss and sampling as well as methodological issues such as the applicability of our measures to other questions and domains.

4.7 Implications

Our methodology combines a multi-level analysis, which zooms in on critical regions of the presented scenes, with summary measures that provide a great deal of information regarding visual preference and exploratory behavior. We find that, at the top level, older TD children pay more attention to faces than younger TD children. This increasing attention trickles down to the next level, where increases in looking times at both *key* and *non-key* features are found. Again, at the bottom level, increases in looking times towards the *eyes* are found. All things being equal, we would expect to find a similar pattern in the number of transitions. This is exactly what we find: as TD children become older they begin to transition more often between *non-key* features and the *eyes*, and between the *eyes* and the *mouth.* By contrast, the number of transitions between *non-*

stimulus and *stimulus* regions does not change, though the total time in *stimulus* regions increases. This pattern suggests that older TD children scan as frequently between the face and extraneous non-face regions as younger TD children, but that when they look at the face, they stay for longer and when they look away from the face, they stay for less, suggesting the development of either an increasing salience for faces, and, in particular, the *eyes*, or an increased ability to ignore distraction and irrelevant scene details.

Entropy results are not as dependent on the total number of transitions as they are on the balance of scanning between regions. At the bottom level, we find increasing exploration between *eyes* and *mouth* as TD children grow older, suggesting increased monitoring of the canonical face scanning circuit. Similarly, at the mid-level, we find a trend towards increased exploration of all areas of faces. Combined with the top-level finding of no changes in *stimulus* to *non-stimulus* scanning, the results suggest that increases in exploratory behavior in TD become more pronounced when zooming into the more inner circuits of face scanning. Thus, in TD children we see a very stable pattern of results, with attention ramping towards areas typically considered informative from a social and communicative point of view as the children age.

By contrast, at the top level, older ASD children look more at *non-stimulus* regions and less at *stimulus regions* than younger ASD children, suggesting a decreasing saliency for faces. This contrasts with ASD results for more inner regions, where no changes with age in looking times for *key, non-key, mouth*, or *eyes* were found. This implies that the trend of decreasing attention towards faces at the top level is not driven by any particular specific face region. Similarly no changes in age were found for transitions and entropy measures in ASD. This suggests that a possible answer to the

question, "how do scanning strategies change in autism?" is: they *don't*. They don't increase looking at *key* features or the *eyes* and they don't increase their exploration of critical regions of the face.

The compounding influence of atypical exposure is a factor which may contribute to the differences we observe. It is useful to consider the effects found in the laboratory within the ecological scope of the affected child (Klin, Jones, Schultz, & Volkmar, 2003). Given that we have found, at every level of our cascaded analysis, behaviors in autism deviating significantly from the typical trajectory, it is an open question as to how these deviations might shape the atypical social and cognitive environment. We note that decreased looking at particularly relevant facial features would imply a decreased experience for those features in that child's life. It is possible that this decreased experience depresses the typical development of configural or holistic face processing, leading to reduced stimulation in certain higher-level cortical social-affective circuitry, which in turn leads to decreased social motivation, which leads back to reduced experience.

In grounding these cascading effects, it is useful to consider the level one (ground level) circuit of the *eyes* and *mouth* for the younger age group. At this age, we find no differences between ASD and TD children in total number of transitions between regions or in entropy. However, we do find that TD children look more at the mouth than ASD children. One possible explanation for this is that TD children are making as frequent transitions between regions as ASD children, but when they encounter the mouth, they stay longer. This mouth salience hypothesis could be driven in part by differences observed in preferences for elementary features. For example, typical children at two

years of age might attend to the mouth in static images, having built up an expectation for motion from this area. By contrast, in ASD, a preference for high areas of contrast combined with decreased sensitivity for motion (Shic, Chawarska, Lin, & Scassellati, 2007, 2008) might bias the child in ASD towards looking predominantly at unique high contrast areas (e.g. the boundary between the sclera and the face or pupil) and less towards the mouth.

It is also possible that trial-level effects are confounding our interpretation of our summary statistics, which are assumed to be valid at a micro-level of analysis. For example, it is possible that both ASD children and TD children attend to the initial area they look upon for a long period of time, and then saccade away to another area. It may be the case that ASD children are more likely to initially focus upon the eyes, and that TD children are more variable in their choice of initial fixation. In this case, the increased variation of TD children could be viewed as noise that, interestingly, increases the chance of uncovering holistic aspects of faces. This interpretation and the perceptual interpretation are not mutually exclusive.

We should note that the trend of older TD children looking longer at the mouth than ASD children is a result which might be considered unexpected given prior results in dynamic scenes by Klin et al. (2002a). There are several possibilities which might account for this difference. First, the switch from a static face scene to a dynamic social interaction scene is a huge leap in cognitive load, social affective circuitry, and basic perceptual saliency. Second, the individuals in Klin et al.'s study were higher functioning individuals with autism; the individuals in this study were more impaired. Third, it is possible that the developmental effects driving mouth-looking simply haven't

come into fruition at the ages examined in this work. For example, if the learning of language interacts with looking at the mouth in autism (for example in aiding phoneme recognition), then it is possible that individuals who are delayed in the use of language would develop compensatory mechanisms much later than four years of age.

Methodologically, it has been our goal to use measures that apply the least amount of data manipulation to the eye-tracking stream as possible. There are several reasons for this. First, nothing is as comparable, or as easy, as doing nothing to your data. Second, the differential loss from complex eye-tracking measures is something that has been little examined; however, our experience is that commonly used eye-tracking tools, such as fixation analysis, routinely allow one to shape one's data in an arbitrary and invisible manner (Shic et al., 2008a, 2008b) . Nonetheless, the comparison of our results to standard fixation measures is a necessary avenue to be explored.

Our measures work, in part, because we have broken down our investigation into several easy-to-digest pieces. For example, our 2^{nd}-order Markov entropy measure can span a single region which is being saccaded over since it has a history of two prior states. As there are at most three regions under consideration at any point in our analysis, this measure is particularly appropriate. This leads us to an important point: the measures employed for analysis should be matched with the appropriate simplifying solutions.

4.8 Chapter Summary

- We have proposed a simple multi-level experimental methodology for eye-tracking analysis and interpretation that controls the combinatorial explosion of region based analysis.
- We have also discussed new avenues for obtaining dynamic measures on eye-tracking data, including entropy techniques to characterize exploration.
- We have used this methodology to obtain a series of results which, though preliminary, have shown some interesting developmental facets regarding the scanning patterns of children with autism, specifically:
    - At 2 years of age TD toddlers and ASD seem very similar in terms of the measures of attention to the face and exploration of these features.
    - For TD children attention to faces seems stronger at 4 years than at 2 years.
    - For ASD children attention to faces does not seem to be different between the two age groups.
    - Attention to the stimulus in general seems to cascade down, affecting all substructures of the face.
    - Differences in attention seem pervasive, affecting both static measures of looking time and dynamic measures of transitions and explorations.

Chapter 5

Computational Modeling of Visual Attention

In the previous chapter, we saw an example of how region-based modeling can provide insight into the scanning patterns of children with and without autism. This methodology, though powerful, requires the intervention of a human guide to trace the regions that are deemed to be worthy of analysis. While it is a simple task to demarcate regions for a limited set of images, in dynamic scenes, such as when subjects to be eye-tracked are presented movies or allowed real-world interactions, the task of marking regions becomes much more difficult and labor-intensive. An alternative method for grounding the gaze patterns obtained by eye-tracking is through computational modeling.

Computational modeling, rather than explicitly specifying the exact range and extent of every item in the scene, defines algorithms which attempt to link image properties to the locations where subjects devote their attention. These models are computational models of visual attention. There are two broad classes of these models. One type of model tries to predict where in the scene a subject will look. These are generative, or predictive, models of visual attention. The other type of model seeks to tie gaze patterns to abstract (though computationally defined) features. These are descriptive models of visual attention. Though the internals of these models might be very similar, indeed, a descriptive model might serve as the basis for a predictive model and a predictive model may contain within it a plausible descriptive model, the role, intent, and evaluation of these two classes of models can be very different.

Predictive computational models of visual attention are often implemented for practical purposes. For example, predictive models of visual attention are used in order to provide front-end processing for robotic visual systems that are expected to interact naturally with people (e.g. Breazeal and Scasselati (1999), also see Fong et al. (2003)). Descriptive computational models of visual attention are often designed to test theoretical ideas, i.e. to test cognitive or perceptual theories that may provide a more unifying description of the underlying process of attention. For example, computational models have been used to describe the process of attention as a process of deriving maximal information regarding the scene (Raj, Geisler, Frazor, & Bovik, 2005; Renninger, Verghese, & Coughlan, 2007; Najemnik & Geisler, 2005). Descriptive models, being driven primarily by theory, may have wide implications, but often have limited application. Typically, the operating characteristics of their performance are limited to certain problems, such as psychophysical or idealized search tasks, or are restricted to the specific features they are constructed to extract. Predictive models, on the other hand, since they are intended to provide some input for real-life application, often try to cover the span of known basic physiological phenomenon, at least in an idealized sense, though in doing so will typically have to make far more and far larger assumptions.

The success of a predictive model is measured by its ability to predict where a subject will look. A natural performance measure for these models, therefore, is comparison against what human subjects do. This is somewhat of a subtle issue, because, as we will see later, the notion of what constitutes similarity in the case of gaze patterns is not immediately obvious. A descriptive model is measured by utility. The strength of a descriptive model is its ability to provide insight. It is possible to have a strongly

predictive model that does not provide a clear understanding as to what is being looked at, just as it is possible to have a very insightful model that does not provide as good of a predictive capacity, as, say, a heavily optimized learning algorithm that collapses information over a set of arbitrary image patches. Naturally, the distinction between these classes of models of visual attention is not quite so stark. As we have mentioned, it is possible to apply machine learning on top of a good descriptive model, just as it is possible to look at the underlying machinery of a strongly predictive model.

In this chapter we will build a general framework for describing computational models of visual attention. Though it is the case that descriptive models of visual attention need not travel all the way to prediction, since there is some overlap between the models, we will describe the framework for a predictive model of visual attention. In reality, the framework itself will remain agnostic as to the use of these models. In order to better illustrate one critical aspect of these models of visual attention, we will examine in depth one popular and often applied model, that of Itti et al. (1998). Since this model was originally intended for the static world, and our goal is to use computational models of visual attention in the dynamic world, we provide a natural extension to the Itti model that captures motion information.

This chapter provides a foundation for the explorations in the next three chapters: Chapter 6, where we will discuss how we can evaluate the predictive capability of models; Chapter 7, where will demonstrate that the line between predictive and descriptive models is not so rigid, and how predictive models can also serve to be descriptive under certain circumstances; and Chapter 8, where we will discuss how these models can be applied in the descriptive sense.

5.1 A Framework for Computational Models of Visual Attention

Computational models of visual attention take as an input some representation of the visual field, perform some processing internally, and return as an output a location upon which attention should be focused (Figure 5.1). Typically, the internal processing can be broken up into two broad components: feature extraction and gaze computation. Feature extraction consists in forming some abstract representation of the raw incoming visual stream and can be arbitrarily complex, ranging from simple filtering methods to systems that employ a wide range of interactions to model the pathways of the human visual system. Gaze computation consists in using the abstract representations generated by feature extraction to determine the location to which attention should be drawn. In many cases, gaze computation can be further broken up into an attention model and a gaze policy. The attention model converts the features generated by feature extraction into an intermediate representation. Often, this intermediate stage is represented as a saliency map that is proportional to, for every spatiotemporal point in the scene, the likelihood that that point will be fixated. A control strategy, the gaze policy, is then applied to the saliency map to generate a fixation point. This can be as simple as choosing the point associated with the highest salience in the saliency map. More formally, the framework begins with a representation of the spatiotemporal scene $\boldsymbol{I}(\boldsymbol{s},t)$ as a function of some spatial coordinate $\boldsymbol{s}$ and temporal index t. This representation is then decomposed, by feature extraction, into a set of features $\boldsymbol{F}(\boldsymbol{s},t)$ that maps in many-to-one fashion onto the original spatiotemporal coordinate system. Operating over these features, an attentional system converts these features into a saliency map, $S(\boldsymbol{s},t)$. Finally, a gaze policy is applied to the saliency map in order to extract a point, $\boldsymbol{g}(t)$, that

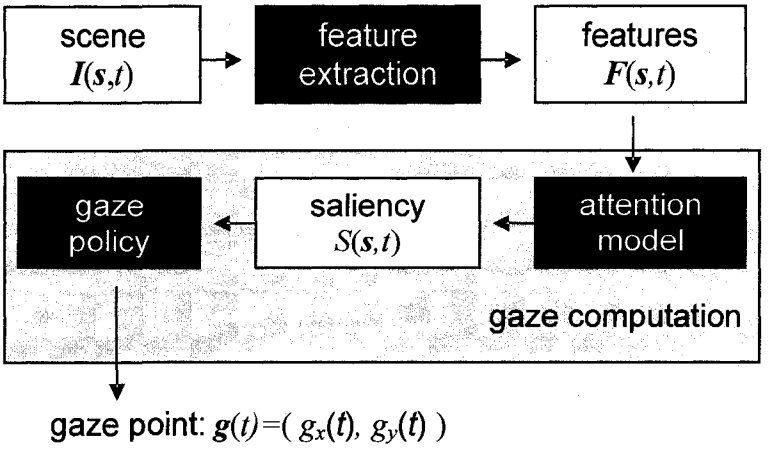


Figure 5.1: Framework for Computational Models of Visual Attention. The spatio-temporal scene $\boldsymbol{I}(s,t)$ is operated upon by feature extraction to provide the features $\boldsymbol{F}(s,t)$. An attention assigns takes the features and computes saliency $\boldsymbol{S}(s,t)$. A gaze policy operates over the saliency to generate the point of gaze $\boldsymbol{g}(t)$.

corresponds to a location that will actually be fixated upon. An example of this process in shown in Figure 5.2. Many influential models of visual attention, such as the biologically-inspired model of Itti (1998) and the psychophysically-driven model of Wolfe (1996), as well as implementations built upon these ideas, such as the context-dependent social behavioral system of Breazeal and Scassellati (1999), obey this formulation.

It is important to note that computational models for visual attention are, by necessity, crude approximations to the human visual attention system and typically operate by identifying, within an incoming visual stream, spatial points of interest. This computational formulation of visual attention is very limiting, in terms of the capabilities and complexities of the biological reality, as many models of visual attention could

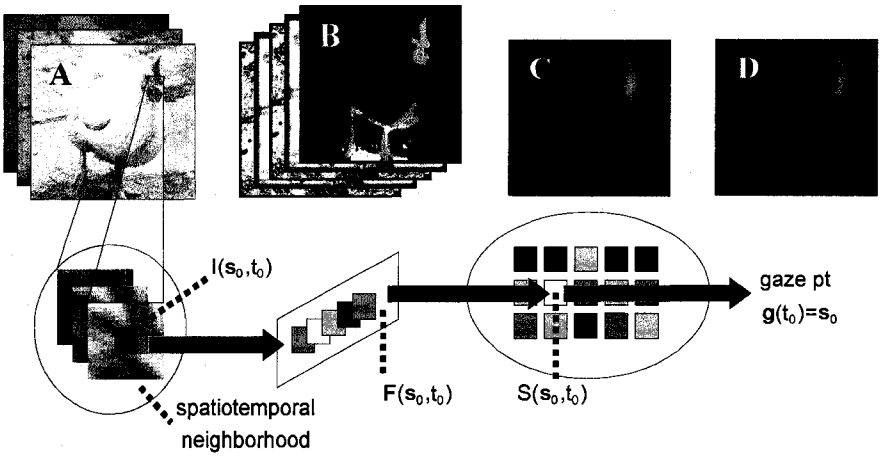


Figure 5.2: Example of a Computational Model of Visual Attention. The original spatiotemporal scene (A) is decomposed into a set of features (B); these features are in turn combined into a salience map (C). From the salience map the location of gaze determined (cross, D).

alternatively be viewed as models of eye fixation or gaze shifting. However, this restricted definition reflects the practical and operational conditions under which our analyses will take place. These models will also serve to (1) reduce the scene to several points of particular interest, thus controlling the combinatorial explosion that results from the consideration of all possible image relationships (Tsotsos, 1988) and to (2) emulate or evaluate the scan-path behavior of human subjects.

5.1.1 Feature Extraction

In computational models of attention, feature extraction is the process of extracting from the input stream abstract representations or key characteristics relevant to the final attentional decision. What exactly comprises the best set of features for guiding visual

attention is an open question, though much progress has been made, especially in areas pertaining to visual search (Wolfe & Horowitz, 2004). Most feature extraction modules, however, choose their attributes based on a combination of biological inspiration and practical concerns. For example, the model of Itti et al. (1998) uses separate channels for image intensity, edge orientation, and color, where each channel is in turn composed of even more elementary channels, such as the "redness" or "brightness" of points.

Note that though the chosen features may be processed early in the visual pathway, their computational formulation or characterization can be arbitrarily simple or complex. For example, by considering an augmented set of features that depend upon previously computed internal variables, we can account for models of selective attention, such as the selective tuning model of Tsotsos et al. (1995), which incorporates bidirectional excitation and inhibition between the feature extraction module and the attention model. This is an important feature, as strictly bottom-up models of visual attention adequately represent neither the true neurophysiological underpinnings of visual attention (Desimone & Duncan, 1995; Posner & Petersen, 1990) nor its computational capabilities and limitations (Tsotsos, 1988).

Our framework does not depend on any specific choice of features; in fact, the utility of our framework depends on the fact that various choices of features may be compared. To guarantee a fair comparison, however, the features to be compared should all be intended to operate over the same type of scenes. In later chapters, we will utilize computational models of attention to examine dynamic environments, such as social situations, and thus we cannot be restricted to static images. Assuming that images from

the visual stream are static suggests that motion is unimportant to visual salience, which is clearly incorrect.

Beyond the requirement that features should acknowledge that there exists a temporal dimension in addition to the spatial dimensions, we do not specify any definite form for features, except that there should exist a set of features associated with every spatial and temporal point of the spatiotemporal scene under analysis. We are then free to choose techniques for feature extraction. The next few sections will include examples of some of the feature sets we will use in later analysis. For simplicity, we will assume that there exist only two spatial dimensions, i.e. our spatiotemporal scenes are 2D-images that change in time. This is an appropriate simplification as our displayed stimuli are movies shown on a computer monitor, and because this approach represents the dominant paradigm in computational work of this nature (e.g. the work of Itti et al. (1998) and Wolfe & Gancarz (1996) use this assumption).

Raw image patch features

Raw image patches (Figure 5.3) are the simplest choice of features associated with some particular spatiotemporal point (s_0,t_0):

$$\boldsymbol{F}(s_0,t_0) = \left\{\boldsymbol{I}(s_0+\delta s, t_0+\delta t)\right\}, \forall \delta s \in \mathbf{N}_s, \delta t \in \mathrm{N}_t \tag{5.1}$$

where N_s is some set of spatial offsets, N_t is some set of temporal offsets, and the two sets together define a spatiotemporal neighborhood in the vicinity of (s_0,t_0). The features that draw attention to a particular point are highly connected to the history of what has transpired near that point. We choose our spatial neighborhood around a spatiotemporal point to be a square centered around the spatial aspect of that particular point, and

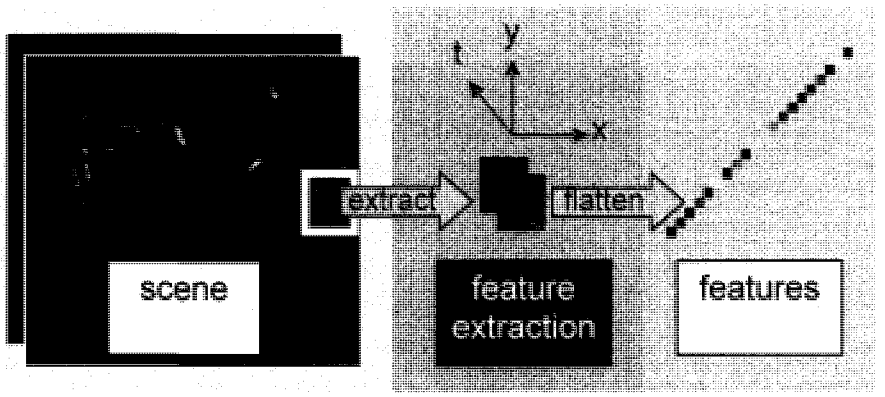


Figure 5.3: Features - Raw image patches. One of the simplest sets of features we can employ is to use all the pixels within a spatiotemporal cube centered at some point in the visual stream as features for that point.

causally from several points backwards in time : $N_s=\{(\delta_{sx},\delta_{sy})\}\ \forall\ \delta_{sx}\in\{-L,L\},\ \delta_{sy}\in\{-L,L\}$ for some characteristic length L.

Gaussian pyramid features

A more satisfying alternative is to build a Gaussian pyramid of the scenes by progressive filtering (Burt & Adelson, 1983) (Figure 5.4). The features corresponding to a point (s_0,t_0) , then, are:

$$F(s_0,t_0)=\{I_i(s_0,t_0+\delta t)\},\forall i\in \mathbf{N}_L,\delta t\in \mathrm{N_t} \tag{5.2}$$

$$I_i(s,t)=I(s,t)*G^i \tag{5.3}$$

where G is a Gaussian filter, and G^i represents i convolutions of G. In other words, the features at a particular point correspond to raw image information at that point, plus the image information of $n+1$ blurred versions of the original image, $N_L=\{0..n\}$, also at that

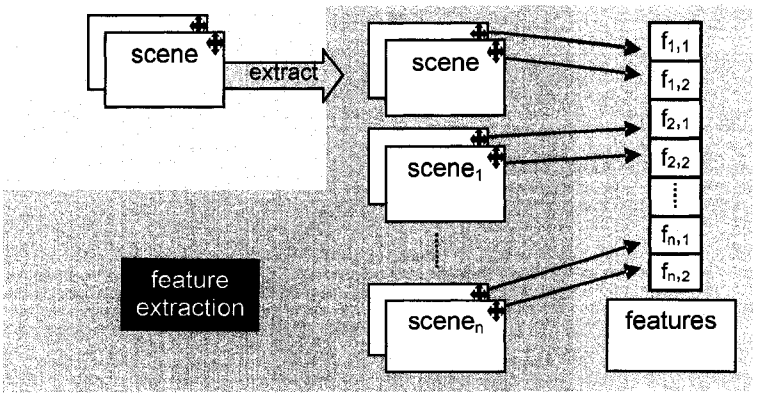


Figure 5.4: Features – Gaussian pyramid. This simple set of features associates every spatiotemporal point (s,t) with the image characteristics $\boldsymbol{I}(s,t)$ and $n+1$ blurred out versions of $\boldsymbol{I}$. In the above figure, $\boldsymbol{G}^i$ is convolved with itself i times. Several time points are also employed in building the features $f_{\lambda,\tau}$, where λ represents the Gaussian level and τ represents the time index.

same point. As was the case for raw image patches, a select set of history is retained in the temporal neighborhood N_t in order to capture the time-varying nature of the scene.

Biologically Inspired Features

Biologically inspired models used for visual attention take the incoming image stream and extract features which approximate the physiology and neurobiology of biological systems. For instance, the model of Itti et al. (1998) generates feature maps $\boldsymbol{F}(s,t)$ corresponding to contrast detection, orientation selectivity, and color sensitivity, using a multi-scale system which includes lateral inhibition. In contrast to the previously mentioned raw image and Gaussian pyramid features, biological inspired features are much more complex, often taking into account the relative physical relationships of wide regions of the entire scene. In Section 5.2 we will examine the model of Itti et al. (1998) in greater depth.

5.1.2 Attention Model

The attention model transforms features associated with a particular spatiotemporal point into a single value that is representative of how likely that point is to be focused upon. In other words, if the original spatiotemporal scene is a color movie with three channels, and this scene is analyzed at the region level to extract D features at every point, the mapping that occurs is $\mathbf{R}^2 \times \mathbf{R}^+ \times \mathbf{R}^3 \rightarrow \mathbf{R}^2 \times \mathbf{R}^+ \times \mathbf{R}^D \rightarrow \mathbf{R}^2 \times \mathbf{R}^+$. The last level in this transformation is the saliency map, a notion originally formulated by Koch and Ullman (1985). We note that though there appears to be some evidence for the coding of an explicit saliency map in the brain (e.g. in the superior colliculus (Kustov & Lee Robinson, 1996); in the lateral geniculate nucleus (Koch, 1984); in V1 (Li, 2002); in V1 and V2 (Lee, Itti, Koch, & Braun, 1999); in the pulvinar (Petersen, Robinson, & Morris, 1987; Robinson & Petersen, 1992); in V4 (Mazer & Gallant, 2003); in the parietal cortex (Gottlieb, Kusunoki, & Goldberg, 1998); general discussion (Treue, 2003)), the question of whether or not saliency maps are actually present physiologically in explicit form has not been answered definitively. Here, we use the saliency map purely as a computational convenience, and where we do not denote saliency as "computational saliency", we hope that it is understood that our work primarily refers to the computational representation of saliency. Computational models of visual attention typically employ saliency maps for computational and organizational reasons and do not necessarily assume a direct biological correlate. Without loss of generality, however, we can employ saliency maps as an intermediate step since any computational model that generates some specific point corresponding to a point of fixation has at least one saliency map that, under some fixed gaze policy, returns the equivalent point. For example, a saliency map that is zero

everywhere except at the point of fixation, where it is positive, will return the correct point under the **arg max** function.

Many different strategies are available for the computation of saliency. Most strategies rely upon the feature integration theory of Treisman and Gelade (1980) which views saliency as the integration of multiple input modality maps, often by linearly weighted summation or nonlinear transfer of linearly weighted summation (Balkenius, Eriksson, & Astrom, 2004; Breazeal & Scassellati, 1999; Itti et al., 1998; Wolfe & Gancarz, 1996). Others view salience in more theoretical terms. For instance, Itti & Baldi (2006) view the salience of spatiotemporal locations in terms of Bayesian "surprise", Torralba (2003) characterizes global contextual factors affecting salience in information-theoretic terms, and Bruce and Tsotsos (2005) use self-information in a neurally plausible circuit to obtain some of the best results regarding overt attention to date. In Chapter 6 we will present another perspective on saliency maps by framing salience as a classification problem on points attended-to by observers and points that are not attended-to.

5.1.3 Gaze Policy

A gaze policy takes the saliency map as input and from it derives the location where attention should be next directed. Formally, if the salience at each point in the saliency map is real-valued, we can simply define this point as:

$$g(t) = \mathbf{arg\,max}_{s}\,(S(s,t)) \tag{5.4}$$

As with the other steps in our framework, the actual implementation of a gaze policy can be more involved, incorporating higher order interactions such as inhibition of return (as in Itti et al., (1998)). Furthermore, the actual action of fixating the eye can involve a change in visual input as the high-resolution fovea rotates to sample the area at a chosen point non-linearly (as in Wolfe & Gancarz, (1996)). Thus there may exist some level of interaction between the gaze policy and the scene input to the system, completing a circuit describing this framework for visual attention.

5.2 The Itti Model

Here we discuss one of the more prominent computational models of visual attention, the model of Itti et al. (1998), as it is a component to many of the studies in later chapters. For simplicity (and hopefully without offense), we will refer to this model of Itti et al. (1998) as "the Itti Model", despite the model having roots at least as far back as Niebur et al. (1995) and reflecting the work of multiple researchers. We should note that, despite our focus on the Itti model, there exist many alternative computational models of visual attention (Balkenius et al., 2004; Breazeal & Scassellati, 1999; Tsotsos et al., 1995, 2005; Wolfe, 1994; Wolfe & Gancarz, 1996; Zaharescu, Rothenstein, & Tsotsos, 2005). These models remain to be explored, though it is a strength of the strategy we will develop in Chapter 6 that a common methodology can be applied to their evaluation.

The Itti Model is a feed-forward bottom-up computational model of visual attention, employing, at its most basic level, decompositions into purely preattentive features. This gives advantages in both speed and transparency. It is a model that is not

only simple but also rigorously and specifically defined, a strong advantage for implementation, extension, and reproducibility of results. It is also possible to download the source code for the Itti model (Itti, 2008), though in our work we implemented the Itti Model in Matlab directly from Itti et al. (1998).

The Itti model extracts the preattentive modalities of color, intensity, and orientation from an image. These modalities are assembled into a multiscale representation using Gaussian and Laplacian pyramids. Within each modality, center-surround operators are applied in order to generate multiscale feature maps. An approximation to lateral inhibition is then employed to transform these multiscale feature maps into conspicuity maps, which represent the saliency of each modality. Finally, conspicuity maps are linearly combined to determine the saliency of the scene. These operations are summarized in Figure 5.5.

The original Itti Model (Itti et al., 1998) did not include a modality for motion. This was rectified by later work (Itti, Dhavale, & Pighin, 2003; Yee & Walther, 2002). However, there seems to be a mismatch between the theoretical concerns of the model and the implementation, as these models do not take into account the direction of motion. The reasons for this discrepancy are not clear. In this work, we use a different formulation for motion saliency (Section 5.2.2) which resulted in better empirical performance. The differences between this formulation (the addition of which will lead to the Extended Itti Model) and previous work are subtle. However, our experience with our own implementation, including use on a humanoid robot, has shown the formulation presented in Section 5.2.2 to be both reasonable and robust.

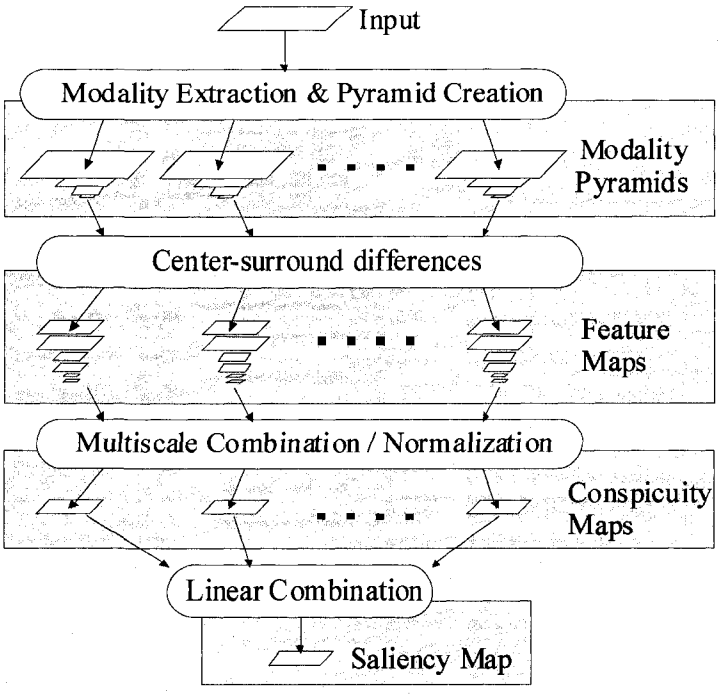


Figure 5.5: Itti Model general architecture (adapted from Itti et al. (1998)). Modalities such as color, intensity, and orientation are extracted and operated on over several stages in order to produce the saliency map associated with the input.

Modality	**Description**
Intensity	Contrasts in luminance; e.g. a small bright area on a larger darker background
Orientation	Pop-out effects based on differences in orientation; e.g. a single diagonal bar in a grid of horizontal bars
Color	Pop-out effects based on color contrasts; e.g. a single red object on a background of green.
Motion	Contrasts in motion; e.g. an object moving to the left as many other objects move to the right

Table 5.1: Description of Modalities in The Extended Itti Model

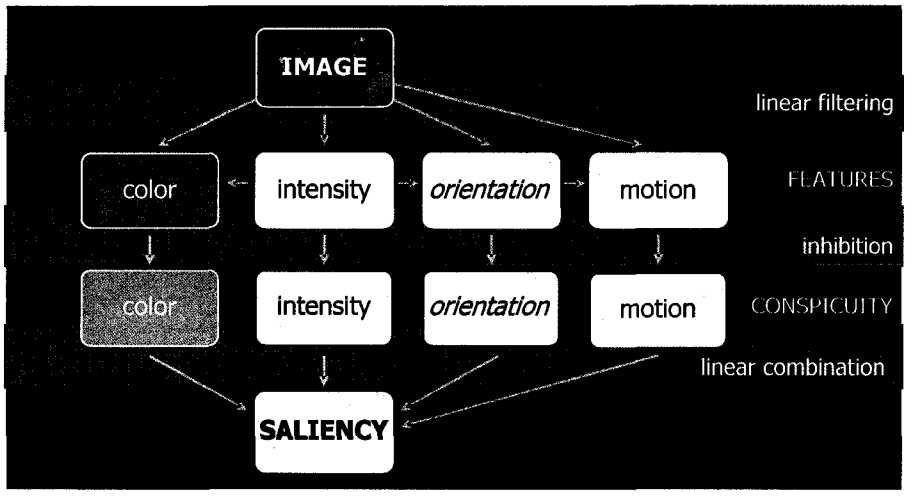


Figure 5.6: Relational diagram of Extended Itti Model. At the feature level, intensity informs orientation, which in turns informs motion. At the conspicuity level, aspects of each modality compete within that modality. The results of conspicuity computation are then funneled into a final saliency.

In the next two sections, for completeness, we will rigorously define the algorithms involved in the Extended Itti Model, which we use for our subsequent analyses. For those who would prefer to skim or skip the more technical descriptions of the Extended Itti Model, we have included an overview in Table 5.1 and a schematic representation in Figure 5.6.

5.2.1 Itti Model for Static Images

We begin by describing the original Itti model, which was intended to operate over static images (motion will be summarized in Section 5.2.2). For additional details, including the rationale for many of the equations and operations, see Itti et al. (1998).

The first stage of the Itti model is to create a multiscale representation of each modality. Given an input image with three color channels, red (r), green (g), and blue

(b), the Itti Model first computes the associated intensity of the image as $I=(r+g+b)/3$. I is then used to create the Gaussian pyramid $I(\sigma)$ and the Laplacian pyramid $L(\sigma)$, where σ is the pyramid scale, in the following filter-subtract-decimate manner (Burt & Adelson, 1983):

$$I^0(n+1) = W * I^0(n) \tag{5.5}$$

$$L(n) = I^0(n) - I^0(n+1) \tag{5.6}$$

$$I(n+1) = \textbf{SUBSAMPLE}[I^0(n+1)] \tag{5.7}$$

with $I^0(0) = I$, the Gaussian filter $W=W_0W_0^T$, $W_0^T=[1/16, 1/4, 3/8, 1/4, 1/16]$, and **SUBSAMPLE** a function which subsamples the input image by a factor of 2. The scales created are $\sigma\in[0..8]$. An example Gaussian pyramid of intensity is shown in Figure 5.7.

The same filter-subtract-decimate method is applied to the individual color channels, r, g, and b, to obtain a multiscale representation of colors, $r(\sigma)$, $g(\sigma)$, and $b(\sigma)$. Normalized color maps at each scale, $r'(\sigma)$, $g'(\sigma)$, and $b'(\sigma)$, are then computed by point-by-point division of color with intensity (points with intensities in $I(\sigma)$ less than $1/10^{th}$ the maximum of $I(\sigma)$ are zeroed). These normalized color maps are combined to yield broadly tuned color channels red (R), green (G), blue (B), and yellow (Y) for each scale (see Figure 5.8):

$$R(\sigma) = r'(\sigma) - (g'(\sigma) + b'(\sigma))/2 \tag{5.8}$$

$$G(\sigma) = g'(\sigma) - (r'(\sigma) + b'(\sigma))/2 \tag{5.9}$$

$$B(\sigma) = b'(\sigma) - (r'(\sigma) + g'(\sigma))/2 \tag{5.10}$$

$$Y(\sigma) = (r'(\sigma) + g'(\sigma))/2 - |r'(\sigma) - g'(\sigma)|/2 \tag{5.11}$$

Figure 5.7: Gaussian pyramid of intensity $I(\sigma)$ with $\sigma \in [1..8]$ (left to right). Each level of the pyramid is computed by downsampling the level above it by a factor of two.

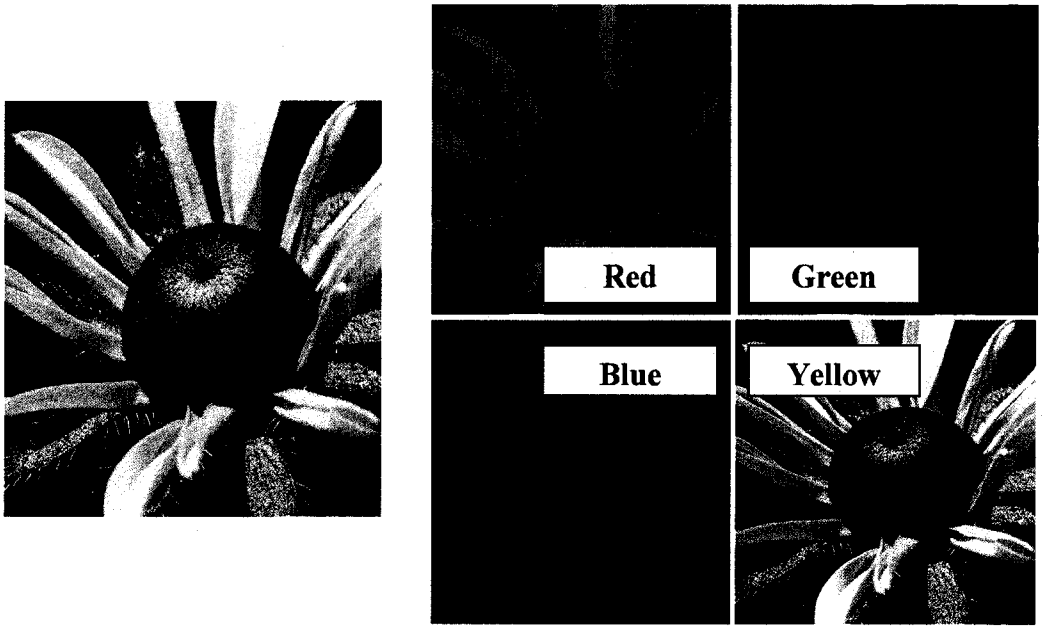


Figure 5.8: Broadly-tuned color maps of the Itti model. The original image is shown to the left, and the component red, green, blue, and yellow color maps are shown to the right. Note that the images are shown on an absolute scale (e.g. the image contains blue, but as the intensity of blue is low, the color map of blue appears dark).

Orientations at multiple scales are computed by taking the real component of spatial Gabor filtering over levels of the Laplacian pyramid (described in (Burt & Adelson, 1983) with alternative notation and slight variation):

$$O_c(\sigma,\theta) = F_s(\theta) * L(\sigma) \tag{5.12}$$

$$O(\sigma,\theta) = \mathcal{R}e\{O_c(\sigma,\theta)\} \tag{5.13}$$

with $F_s(\theta)$ the coarse Gabor filter at orientation $\theta=\pi N / 4$, $N\in[0..3]$, defined in 2D for a given point at (x,y):

$$F_{s;x,y}(\theta) = W_{x-x_0,y-y_0} e^{i\frac{\pi}{2}(x\cos\theta+y\sin\theta)} \tag{5.14}$$

where W is the Gaussian filter used in (1), with x_0 and y_0 chosen to appropriately center W ($x_0 = y_0 = -2$ in our case) (see Figure 5.9).

From these multi-scale representations of intensity, color, and orientation, feature maps are derived. Feature maps are created with the aid of a center-surround difference operator Θ. For a given multi-scale modality X, $X(c)\ \Theta\ X(s)$ interpolates the image with lower resolution to the resolution of the finer image, and then subtracts point-by-point. The interpolation is accomplished through the inverse application of equations 5.5 and 5.7. For all modalities, the center scales are $c\in\{2,3,4\}$ and the surround scales are $s=c+\delta$, $\delta\in\{3,4\}$.

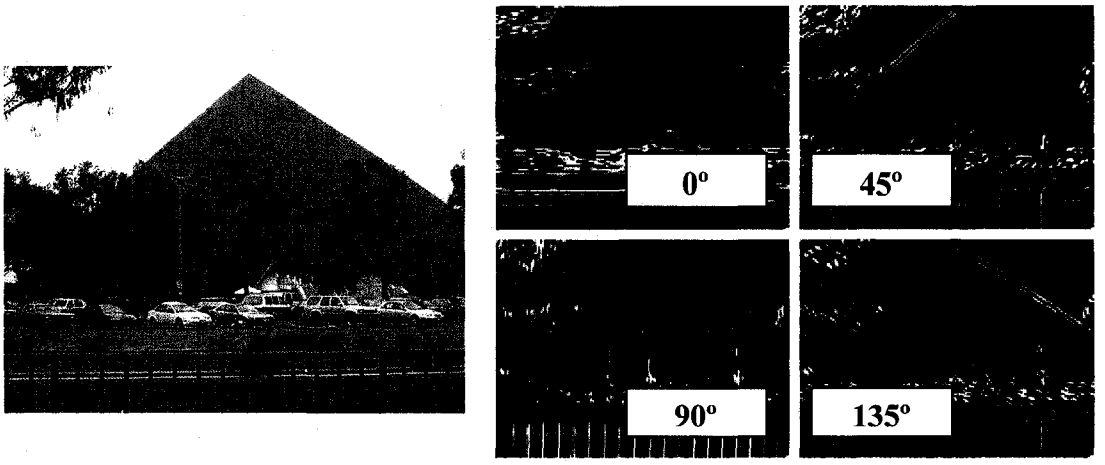


Figure 5.9: Orientation selection of the Itti model. The original image is shown on the left and the result of filtering with Gabor filters of orientation 0° (horizontal), 45° (bottom left to top right), 90° (vertical), and 135° (bottom right to top left) is shown on the right. Note the pattern of the bottom gate at 0° and 90° and the emphasis on the left and right pyramid side at 45° and 135°.

The intensity feature maps $\mathcal{I}(c,s)$ are straightforward (Figure 5.10):

$$\mathcal{I}(c,s) = | I(c) \ominus I(s) | \tag{5.15}$$

The color feature maps are slightly reordered to emulate color double-opponency for red-green $\mathcal{RG}(c,s)$, and blue-yellow $\mathcal{BY}(c,s)$ (Figure 5.11):

$$\mathcal{RG}(c,s) = | (R(c) - G(c)) \ominus (G(s) - R(s)) | \tag{5.16}$$

$$\mathcal{BY}(c,s) = | (B(c) - Y(c)) \ominus (Y(s) - B(s)) | \tag{5.17}$$

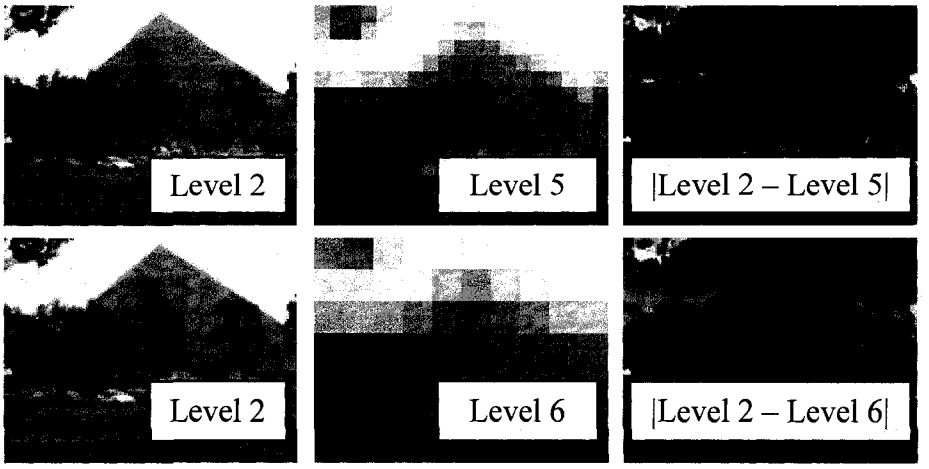


Figure 5.10: Intensity feature maps $\mathcal{I}(c,s)$. The left column is the center image, the middle column is the surround image, and the right image is the intensity feature map. The top row shows $\mathcal{I}(2,5)$ and the bottom row shows $\mathcal{I}(2,6)$.

Figure 5.11: Color double-opponency maps for red-green $\mathcal{RG}(c,s)$ (left), and blue-yellow $\mathcal{BY}(c,s)$ (right). Maps are created from the image in Figure 5.8 by center-surround calculations on selected color maps for a center at level 2 and a surround at level 5.

Finally, the orientation feature maps $\mathcal{O}(c,s,\theta)$ are separately coded for each orientation Θ (Figure 5.12):

$$\mathcal{O}(c,s,\theta) = | \mathcal{O}(c,\theta) \ominus \mathcal{O}(s,\theta) | \qquad (5.18)$$

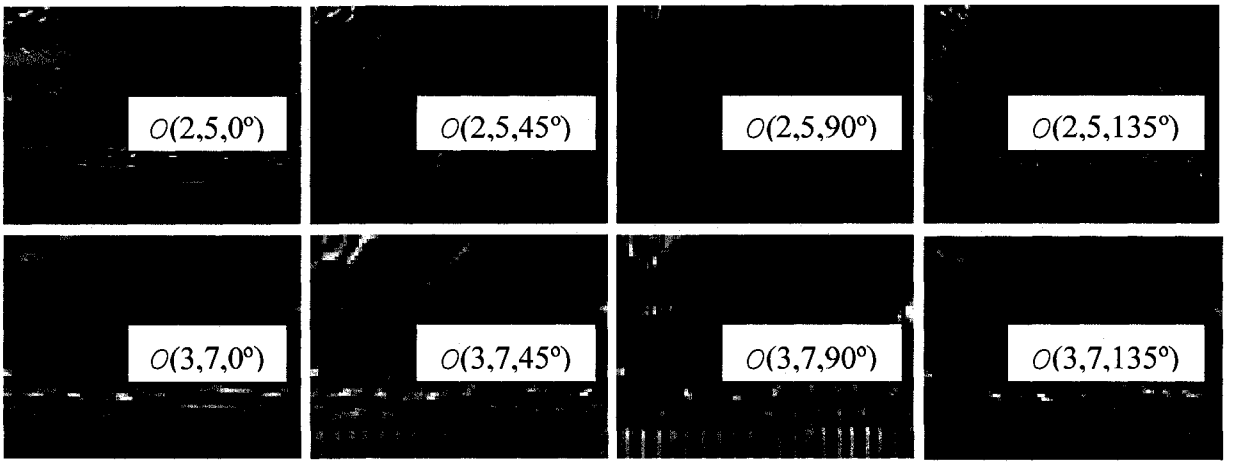


Figure 5.12: Orientation feature maps $O(c,s,\theta)$ created as center surrounds for different orientations. The top row shows the orientation feature maps for c=2, s=5, and $\theta \in \{0°, 45°, 90°, 135°\}$, and the bottom row shows the orientation feature maps for c=3, s=7.

Feature maps for each modality are then combined into conspicuity maps. Conspicuity maps represent the salience of the modality as a whole. This is mediated through a normalization operator, $\mathcal{N}$, and a cross-scale addition operator, $\oplus$. The normalization operator $\mathcal{N}(\mathcal{M})$ returns a rescaled version of map $\mathcal{M}$, approximating lateral inhibition (Itti & Koch, 1999), by first linearly scaling $\mathcal{M}$ into a fixed range $[0,M]$, then multiplying the map by $(M-m)^2$, where m is the average of all local maxima in $\mathcal{M}$ except one point where the value is M. In our work, local maxima were locations with values greater than all eight neighbors. The cross-scale addition operator, $\oplus$, expands or reduces maps to scale 4 and then adds point-by-point.

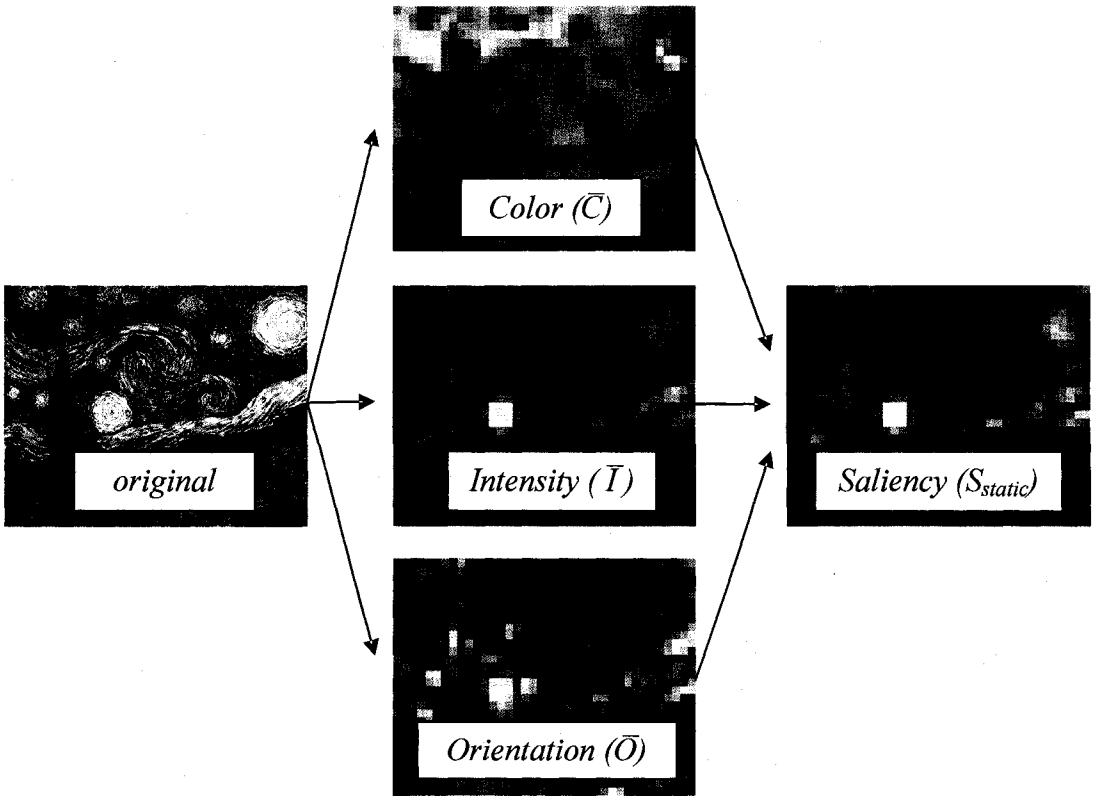


Figure 5.13: Computational of final saliency map (S_{static}). The original image is decomposed into feature maps which are in turn assembled into conspicuity maps of color ($\bar{C}$), intensity ($\bar{I}$), and orientation ($\bar{C}$). These conspicuity maps are then summed together to obtain the final saliency map (S_{static}).

The intensity conspicuity map $\bar{I}$, color conspicuity map $\bar{C}$, and orientation conspicuity maps $\bar{O}$ are then defined (see Figure 5.13):

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{I}(c,s)) \tag{5.19}$$

$$\bar{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\, \mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s)) \,] \tag{5.20}$$

$$\overline{O} = \sum_{\theta \in \{0, \frac{\pi}{4}, \pi, \frac{3\pi}{4}\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \right) \tag{5.21}$$

Finally, the various conspicuity maps are combined under normalization in order to generate the final saliency map, S_{static} (Figure 5.13). Note that we have designated the saliency map to be static to distinguish it from the saliency map generated in the next section which would be more applicable for dynamic scenes.

$$S_{static} = \frac{1}{3}\left(N(\overline{I}) + N(\overline{C}) + N(\overline{O})\right) \tag{5.22}$$

5.2.2 Extended Itti Model for Dynamic Scenes

This Itti Model, originally intended to operate over static images, is not quite an appropriate fit for analyzing the dynamic scenes which are the stimuli used in many of our experiments. For this reason, we extend the model with the ability to characterize motion. This extension derives from Shic & Scassellati (2007).

Simple image difference detection (by computing the absolute difference between adjacent frames, as is done in Niebur and Koch (1996)) is insufficient as a basis for a motion modality, as it fails to highlight known pop-out effects (Figure 5.14 right) (Wolfe & Horowitz, 2004). Similarly, employing optical flow (see Beauchemin & Barron (1995) for a review) as a basis for motion salience typically involves a retrospective interpretation of the optical flow field, a paradigm that does not fit neatly into the feedforward framework. Optical flow techniques which could be easily adapted, such as

Heeger's work with hierarchical spatiotemporal Gabor filters (1988), are computational expensive as they incorporate numerical optimization at each image location.

We employ a compromise approach as a basis for computing motion saliency, a variation of time-varying edge detection (as recounted in Jain et al. (1995)). The time-varying "edginess" of a point, $E_{s,t}$, is computed as the product of the spatial and temporal derivatives:

$$E_{s,t}(x,y,t) = D_s I(x,y,t) \cdot D_t I(x,y,t) \tag{5.23}$$

where I is the intensity of an image at the spatial coordinates, x and y, and at the temporal coordinate, t, and s is some spatial direction $s=s(x,y)$. In our work, we approximate the spatial derivative with the imaginary component of the Gabor-filtered image obtained during the basic Itti Model extraction since the imaginary component of the steerable filter functions as a spatial edge detector (see Greenspan et al., 1994). We obtain the temporal derivative from image differencing after temporal filtering. Note that this technique can only provide the combined magnitude of motion and intensity and not the magnitude of stimuli motion alone. This flaw, however, is mitigated by the multi-scale aspect of the Itti Model. Our motion extension is very much in the style of the Itti model as it is (1) integrated in a fashion similar to that of the orientation modality and does not break away from the original model's methodology or framework, (2) computationally quick and easy to implement, and (3) capable of describing a wide range of pop-out motion phenomena. The relational diagram for the full Extended Itti Model is shown in Figure 5.6.

We begin by extending the original Itti Model (Itti et al., 1998) equations in time (e.g. the intensity modality $I(\sigma)$ becomes $I(t,\sigma)$, the red-green feature map $\mathcal{RG}(c,s)$ becomes $\mathcal{RG}(t,c,s)$, etc.) Working purely with image intensities, under the assumption that motion is largely color-blind, for N frames $I(t,\sigma)$, $t \in [1..N]$, we obtain motion feature maps in the following manner:

1) Compute the N-th order backwards-difference approximation to the temporal derivative, $\mathcal{M}_t(t,\sigma)$ (the first order is shown here; for higher order approximations see Khan & Ohba, 1999):

$$\mathcal{M}_t(t,\sigma) = I(t,\sigma) - I(t-\Delta,\sigma) \tag{5.24}$$

2) Compute the spatial derivative, $\mathcal{M}_s(t,\sigma,\theta)$, from gradients extracted during orientation computation:

$$\mathcal{M}_s(t,\sigma,\theta) = Im\{O_c(t,\sigma,\theta)\} \tag{5.25}$$

3) Compute the motion feature map $\mathcal{M}(t,\sigma,\theta)$ as the product of $\mathcal{M}_s(t,\sigma,\theta)$ and $\mathcal{M}_t(t,\sigma)$:

$$\mathcal{M}(t,\sigma,\theta) = \mathcal{M}_s(t,\sigma,\theta) \cdot \mathcal{M}_t(t,\sigma) \tag{5.26}$$

The motion conspicuity map is derived directly from the above algorithm, using the normalization operator $\mathcal{N}$, and a cross-scale addition operator, $\oplus$, as defined in Itti et al (1998), to emulate the effects of lateral inhibition:

1) Compute the direction of motion for each orientation to obtain positive and negative directional features. The positive directional feature $\mathcal{M}_+(t,\sigma,\theta)$ is defined as $\sqrt{\mathcal{M}(t,\sigma,\theta)}$ at locations where $\mathcal{M}(t,\sigma,\theta)$ is positive, and 0 otherwise. Similarly, the negative directional feature $\mathcal{M}_-(t,\sigma,\theta)$ is defined as $\sqrt{-\mathcal{M}(t,\sigma,\theta)}$ at locations where $\mathcal{M}(t,\sigma,\theta)$ is negative, and 0 otherwise.

2) Compute the directional contribution to motion conspicuity, $\mathcal{M}_d(t,\sigma,\theta)$ by allowing positive and negative directional motion features to compete locally:

$$\mathcal{M}_d(t,\sigma,\theta) = \mathcal{N}(\mathcal{M}_+(t,\sigma,\theta)) \oplus \mathcal{N}(\mathcal{M}_-(t,\sigma,\theta)) \tag{5.27}$$

This accounts for popout phenomena such as that shown in Figure 5.14.

3) Compute the across-scale contribution for each orientation, $\mathcal{M}_o(t,\theta)$.

$$\mathcal{M}_o(t,\theta) = \bigoplus_{\sigma=0}^{8} \mathcal{N}(\mathcal{M}_d(t,\sigma,\theta)) \tag{5.28}$$

This is equivalent to saying that all scales at a particular orientation compete with one another.

4) Compute the conspicuity map for motion, $\overline{M}(t)$, by combining across all orientations:

$$\overline{M}(t) = \sum_{\theta\in\{0,\frac{\pi}{4},\pi,\frac{3\pi}{4}\}} \mathcal{N}(\mathcal{M}_o(t,\theta)) \tag{5.29}$$

Motion conspicuity is then added, as an additional modality, to obtain the final saliency map S:

$$S = \frac{1}{4}\left(\mathcal{N}(\bar{I}) + \mathcal{N}(\overline{C}) + \mathcal{N}(\overline{O}) + \mathcal{N}(\overline{M})\right) \tag{5.30}$$

replacing the basic Itti Model equation for saliency in Equation 5.22 (which only neglects a term for motion).

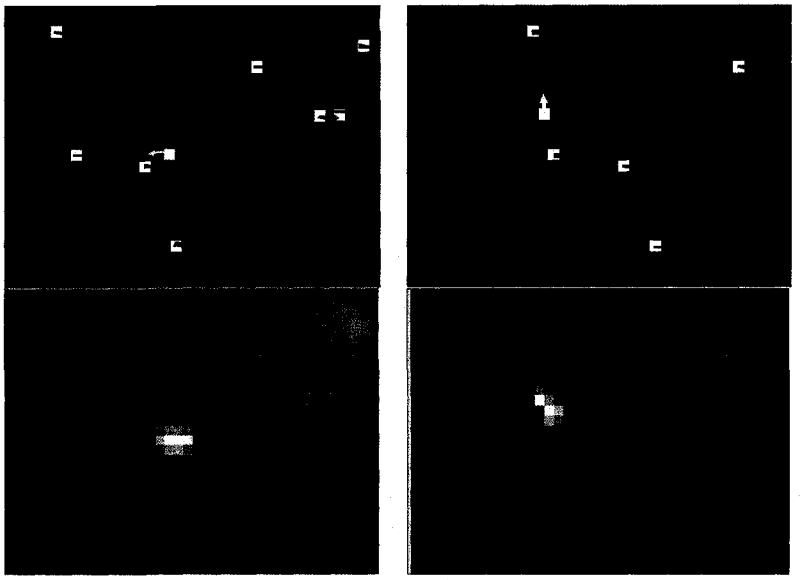


Figure 5.14: Motion pop-out stimuli composed of boxes (top panes) and associated final motion conspicuity map (bottom panes). The left figures represent directional competition, with a single stimuli moving leftwards in a field of rightward moving distractors. The right figures represent cross-orientation competition, with a single upwards moving stimuli popping-out among rightward moving distractors. The arrows in the top panes are for illustrative purposes only, and did not appear in the actual stimulus.

Given the framework and models that we have discussed in this chapter, we will now turn towards applying these models to human gaze data. In the next three chapters we will employ computational models in both the predictive sense and in the evaluative sense. We will find that, though the choice of a particular computational model is not without its subtleties, there exists a wide array of uses for these models in both comparative analyses and exploratory investigation.

5.3 Chapter Summary

- We have presented a general framework for computational models of visual attention where scenes are decomposed into features, features are translated into a saliency map, and a gaze point is extracted from the saliency map.
- We have discussed the feature extraction phase at length and offered several examples of features: raw image patch features, Gaussian pyramid features, and biologically-inspired features.
- We have discussed one biologically-inspired set of features, that of Itti et al. (1998), and have provided a detailed account of its mechanisms.
- We have developed a novel motion extension to the Itti model that captures basic motion pop-out effects.

Chapter 6

Comparing Predictive Models of Visual Attention

Several methods could be used in order to compare predictive models of visual attention to human gaze trajectories. First, we could assemble a series of tasks with associated visual scenes and compare the performance on these tasks. For example, we could use the region-based modeling approach of Chapter 4 and determine performance on summary scores in those regions (e.g. the total amount of time spent looking at each region) for both human observers and the computational models. This is a valid solution. However, if the region-based analysis was geared towards describing high-level effects, such as a preference for smiles, the computational model would most likely have to be formulated in terms of these same high-level processes. In addition, it is possible that unless the summary scores were cut very fine, they would miss much of the time-varying dynamic of human scan patterns. Also, such an approach would not readily lend itself towards describing individual variation unless the models explicitly built in some learning or adaptability.

A second possibility is that we could use a more computational paradigm to evaluate predictive models. In this chapter we will discuss how one can compare predictive models of visual attention through an automated process. First, we will define what it means for two gaze patterns to be considered different. This implies that we should search for a suitable metric, one that is grounded somehow in the predictive capability of each model. However, even with such a metric, we will still suffer from the "baseline" problem. This problem is loosely described as follows: given that you have

some computational model that aims to predict the location of gaze, and given that this model is controlled by any number of free parameters, how can we compare one model against another when two models are tuned differently? Put in another way, how can we compare models meant to operate in different domains, when the task presented might be arbitrarily close to one of the models just by chance? The answer is to bring the models out of their baseline, and to apply some level of optimization to all models such that the test of predictive ability is a fair one. This is the basis for our classification strategy for computational saliency. Finally, how does one obtain a good measure of how well a model fits? The point of regard of a single individual at any given time is confined, by its nature, to a very small portion of the scene. To obtain a smoother representation of "goodness of fit", we will apply gaze indexing on the rank ordering of the intermediate representation of salience, the saliency map. We will conclude with an experimental comparison of the models implied by the features described in Chapter 5. The work in this chapter has been derived from our previously published work (Shic & Scassellati, 2007).

6.1 Metrics for Modeling Visual Attention in Dynamic Environments

To compare models against human subjects, we need to define some metric over which some notion of similarity can be made. An obvious choice for such a metric is gaze fixation distance. We can say that a particular gaze process $\mathbf{G_a}$ is close to another gaze process $\mathbf{G_b}$ if the points of fixation chosen by $\mathbf{G_a}$ and $\mathbf{G_b}$ are spatially close for all points in time. However, a major problem with distance measures is highlighted in Figure 6.1.

In the case of Figure 6.1 left, if some model picks point A and another model picks point B, we could safely say that these two models are dissimilar. Conversely, if one model picks A and another model picks C, we could say that the models are similar. In this case, a distance metric based on distance between fixations makes sense. In the case of Figure 6.1 right, if one model picks A' and another model picks B', we can still say that these two models are similar. However, if one model picks A' and another model picks C', the distance is much greater than A'-B'. However, the underlying image content at points A' and C' are very similar. In this case, using a fixation distance metric does not make sense. By employing distance metrics between points of fixation, we ignore the underlying substrate of visual attention: that of the scene itself. Essentially, employing distance as the sole measure of similarity results in questionable results since gaze patterns are dependent on the underlying scene. Note that this is true whether we employ distance directly or use some nonlinear variant that is dependent upon distance, such as overlap of Gaussians centered at fixation points.

An alternative to using distances for comparison is to use some index of saliency as the measure. This is the method employed in both Parkhurst et al. (2002) and Ouerhani et al (2004). Notably, both groups use the locations that human subjects fixate upon to index into the saliency map, and show that the saliency at the locations attended to by humans is greater than what would be expected by a random process. Since saliency is assembled from features, and since features change in a time-varying fashion, the technique of collapsing eye movements across time, as done by Ouerhani, is not applicable to our environment. For instance, if, on the right image of Figure 6.1, we were to show only one face, and after some short time, cover that face and display the other

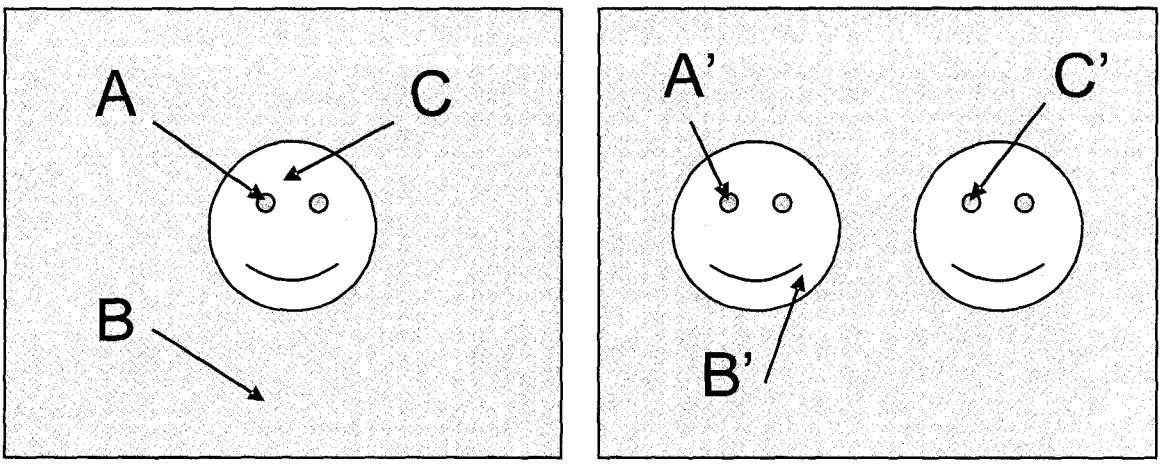


Figure 6.1: Problems with Fixation Distance Metrics for Measuring Similarity. *Left:* In a Euclidean sense, A is close to C, and B is not close to either. Here Euclidean measure of similarity makes sense. *Right:* In a Euclidean sense, A' and B' are closer than A' and C'. However, both A' and C' isolate the eye. They are dissimilar in a Euclidean sense, but similar in terms scene content.

face for a time period equal to the first face display, and if a process $\mathbf{G_a}$ focused on faces in both situations, but a process $\mathbf{G_b}$ focused on the conjugate empty space in the same situations, we would have identically collapsed probability functions, but a very different underlying gaze strategy.

Another alternative is to aggregate the looking points of a large number of human subjects for each point in time. On static images it is easy to obtain 10 seconds of looking time at 60 eye recordings a second for a total of 600 eye recordings over an image for a single individual. For time-varying images, however, we require either a large number of subjects or a set of strong assumptions about the probability fields associated with each eye fixation (such as a Gaussian region centered about each gaze fixation), to obtain the same level of sampling. Using a large sample of individuals, of

course, does provide a great deal of information. However, in the interest of a generalized computational framework for visual attention, we desire a technique that, while still being able to benefit from multiple sources of data, is not completely dependent on the sampling size of human data.

Finally, for non-biologically-inspired models, saliency is not necessarily a cleanly defined concept. Since we want to compare models to models as well as models to humans, it is in our interest to develop some strategy that makes saliency somehow comparable across various selections of features. The method we employ in this work is to define distance at the feature-level. That is, we say that two spatiotemporal locations are "close" if their underlying features are close. The particular implementation of this distance measure is the subject of the next section.

6.2 A Classification Strategy for Computational Saliency

We desire a method for computing saliency from features (ignoring task knowledge and other top-down effects which definitely play a role in biological visual saliency, a point to which we will return after our experiments in Section 6.3). The method that we employ in this work is to divide spatiotemporal scenes into two classes: (1) locations attended-to and (2) locations *not* attended-to. We define saliency as some function that is related to the probability that a particular location, based solely on its associated features, is likely to be fixated upon by a human observer. By defining saliency in this manner we achieve several goals: (1) we obtain a mapping from features to saliency that corresponds to a structured and intuitive measure of distance in feature space; (2) we obtain a method

that makes saliencies for different choices of features comparable, since they represent an underlying likelihood of fixation; and (3) since features are translated directly into saliencies, which represent probabilities, we do not need to optimize an individual model to match a human's gaze pattern – such an effect is incorporated implicitly in the mapping. In the following subsections we provide the mathematical basis for this classification strategy for saliency.

6.2.1 Bayesian Classification Strategy for Attention

We know that, for some feature vector $\boldsymbol{f}$ and class c_i:

$$p(c_i \mid \boldsymbol{f}) = \frac{p(\boldsymbol{f} \mid c_i)p(c_i)}{p(\boldsymbol{f})} \tag{6.1}$$

If we were to use a Bayesian classifier, we would, for two classes c_0 = attended-to and c_1 = not attended-to = $\neg c_0$, choose class c_0 if $p(c_0 \mid \boldsymbol{f}) > \Theta\, p(c_1 \mid \boldsymbol{f})$ for some threshold Θ, and would choose class c_1 otherwise. We could thus define saliency to be:

$$\varphi(\boldsymbol{f}) = \frac{p(\boldsymbol{f} \mid c_0)}{p(\boldsymbol{f} \mid \neg c_0)} \tag{6.2}$$

However, φ can be arbitrarily large, due to the term in the denominator. More problematic is that $p(\boldsymbol{f} \mid c)$ must be estimated. This tends to be quite difficult in high dimensional spaces, and, even in low dimensions, may require more complicated approximation techniques. Note that this formulation is similar to other findings which

take a Bayesian approach towards aligning scene features with points of regard (Itti & Baldi, 2006; Torralba, 2003).

6.2.2 Fisher's Linear Discrimiant Strategy

Though useful as an intuitive conceptualization of the visual attention process, it is not necessary to explicitly form a probability map representing the likelihood of attending to each spatiotemporal location. Attention is directed towards some "interesting" point. In some ways, it does not matter if the function governing the decision to attend to some location is twice or three times the value of some other, less likely to be attended-to, point, only that it be greater. For this reason we can relax the ideal that saliency should correlate directly with a probability, and use forms of dimensionality reduction to aid in the computation of salience. Many dimensionality reduction schemes exist, with varying abilities to adapt to non-linear relationships, and with varying levels of biologically plausibility. The method that we employ here is one of the oldest, and simplest, techniques: Fisher's linear discriminant (Fisher, 1936; Duda et al., 2000).

By using this model, we do not presuppose the existence of any biological or psychophysical effect, and, furthermore, only need to specify that we expect some difference exists between the locations that are attended to and the locations that are not. With the two classes, c_0 and c_1, corresponding to points in the spatial temporal scene where gaze is fixated and points where gaze is not fixated, respectively, the Fisher criterion function $J(w)$ is:

$$J(w) = \frac{w^t S_B w}{w^t S_W w} \tag{6.3}$$

where w is a weight matrix, S_B the between class scatter matrix (equivalent to the outer product of the difference between the class means), and S_W the within class scatter matrix (proportional to the sample covariance matrix; see Duda et al., 2000). Linear discriminant analysis seeks to find the weight w maximizing $J(w)$. Intuitively, this corresponds to finding a projection that maximizes the ratio between the difference between classes (S_B) and the variability of the data in those classes (S_W). This optimization yields the solution for w:

$$w = S_W^{-1}(m_1 - m_2) \tag{6.4}$$

where

$$m_i = \frac{1}{|c_i|}\sum_{x \in C_i} x \tag{6.5}$$

and

$$S_W = S_1 + S_2 \tag{6.6}$$

with

$$S_i = \sum_{x \in C_i}(x - m_i)(x - m_i)^t = \left(|c_i| - 1\right)\Sigma_i = k_i \Sigma_i \tag{6.7}$$

(Duda et al., 2000). The projection matrix w is used to project a location's features to one dimension, and it is this projection that serves to approximate saliency.

Of particular concern, however, is the fact that there is a large asymmetry in the sizes of the populations of classes. At any particular point in time, there are a large number of regions corresponding to points *not* fixated upon, but only one region corresponding to that point upon which gaze *is* fixated. If we were to take equation 6.7 verbatim, then we would end up with a projection predominantly shaped by the covariance of patches that gaze is not fixated upon, and this, in turn, would be similar to the general properties of the spatiotemporal scenes in the data set we choose. Thus the hypothesized difference in covariance structure between gaze-fixated points and non-gaze-fixated points would tend to be washed away. To compensate for the disparity in data sampling we instead assume that the factors k_i in equation 6.7 are equal across classes. This assumption leads to greater discrimination ability between fixated and non-fixated points, as measured empirically.

6.2.3 Rank Ordering

The measure from equation 6.7, a value for saliency at every point, is a projection not in metric proportion. For example, if application of our weight matrix to some map of features were to yield a particular value at one point, and half that value at a second point, we should not interpret this to imply that the second point was half as likely to be focused upon. We could reasonably assume, however, that the second point was less likely to be focused upon. For this reason, instead of using the *value* of saliency directly, we examine our samples in terms of the *ordering* of their saliency at a given point in time. In other

words, we assume that, for a given time t, the saliency at any spatial location s can be compared with the saliency of other spatial locations via ranking. We do not assume that saliency computations at different points of time can be directly compared, as this would imply that our saliency measure in some way represented a global metric with global implications rather than a local metric over local features.

It is important to note that many alternative strategies exist for normalizing saliency. We could, for instance, require the saliency map span a range of values from 0 to 1, or that the energy of the saliency map be normalized. These normalization strategies require different sets of assumptions. Our choice of attention model will greatly impact these relationships. For instance, the Bayesian strategy presented in Section 6.2.1, formulated as a ratio, and the Fisher discriminant strategy presented in Section 6.2.2, formulated as a projection, result in two very different distributions. Since maintaining comparability despite changes in the underlying attention model or feature extraction process is one of the goals of this work, we employ a final measure that is independent of monotonic transformations on saliency.

6.3 Evaluation of the Itti Model as a Predictive Model

In regards to biological and theoretical models of attention, a large body of work exists (for specific issues relevant to this work, see Itti, Rees, & Tsotsos (2005)). In regards specifically to the Itti Model several quantitative analysis have been accomplished. Parkhurst, Law, and Niebur (2002) show that the saliency maps of images, as computed by the Itti model, is higher in locations fixated upon by human subjects than would have

been expected by chance alone. Ouerhani et al. (2004) show that the saliency maps generated by the same computational attention model are correlated to approximate probability density maps of humans. In Itti et al. (2003), temporal flicker and a Reichardt model for motion are added to the Itti model, allowing for analysis of dynamic scenes. Using this augmented set of features, Itti (2005) shows that, in short movie clips, the salience of this augmented model is higher at the target of human saccades and that the motion and temporal components of the model are the strongest predictors of these saccades. Most recently, Carmi and Itti (2006) show that shortly after jump-cuts, when bottom-up influences are presumably strongest, these dynamic components have even greater ties to human saccades.

The Itti model and the interpretations of its results are not uncontroversial. Turano et al. (2003) shows that the gaze locations predicted by the static Itti et al. (1998) model are no better than random, in direct contrast to the Parkhurst et al. (2002) results. This experiment, however, uses a different measure of performance, comparing the unique model predicted gaze location to the human gaze locations, and also takes a static model and applies it to a dynamic environment. Tatler et al. (2005) employ an alternative set of elementary features as well as a different set of measures for performance to provide an alternative interpretation of the results of Parkhurst et al. (2002). Draper and Lionelle (2003)(2005) show that the iLab Neuromorphic Vision Toolkit (Itti, 2008), an implementation of the Itti model, is not scale or rotation invariant, thus questioning the appropriateness of using the Itti model as the basis of computational object recognition systems. Finally, Henderson et al. (2007) show that the Itti model cannot account for human behavior during search tasks.

Though there are similarities between our study and the aforementioned work, noticeable differences exist. First, our work employs a new metric for measuring the distance between the gaze patterns of models and individuals based on classification performance and dimensionality reduction. This contrasts with studies which use Euclidean-based measures and is more similar, but not equivalent to, those studies that employ similarity based measures. Second, our work is not compatible with previous works which operate over static images (Ouerhani et al (2004), Parkhurst et al. (2002), and subsequent discussions). The addition of a temporal component complicates analysis: human scan trajectories cannot be collapsed across the time dimension when the underlying substrate of attention, the visual scene, is time-varying. Third, most studies choose default "mixing parameters" for the contribution of, say, color over intensity, in the final calculation of the salience map. In reality, the actual contribution of different modalities is likely to be neither strictly linear nor strictly equivalent. Computational models of attention can benefit from some optimization of parameters to match human gaze patterns, thus revealing statistics regarding the capacity of a model versus its default performance. In our work, optimization occurs as a byproduct of viewing gaze selection as a classification and dimensionality reduction problem, as we saw in Section 6.2.

In this section, we will compare different computational models of visual attention against human subjects. However, since we are comparing multiple models, we must also control for multiple sources of variation, such as the inherent dimensionality and spatiotemporal extent of the underlying features. By comparing a wide range of parameters on our computational models, and by choosing good controls for our human subjects, it is hoped that these sources of variation can be controlled.

6.3.1 Subjects and Data

The human subjects in this experiment consist of 10 individuals drawn from a population of adolescents and young adults that are intended to serve as age and verbal-IQ matched controls for a different study, one which compares these controls versus individuals with autism (Klin et al., 2002a). While this group is predominantly considered normal, some of the individuals of the population fall in a range that labels them as mildly mentally retarded. It is our intent to conduct this experiment over subjects that are slightly varied in mental capability, as we do not expect our technique to hinge on a notion of a "typical" human subject.

The gaze patterns for these human subjects are obtained via a head mounted eye-tracker (ISCAN Inc, Burlington, Massachusetts) under controlled conditions as the subjects watch two different, approximately 1 minute long, clips of the 1966 black and white movie "Who's Afraid of Virgnia Woolf". The eye tracker employs dark pupil-corneal reflection video-oculography and has accuracy within $\pm 0.3^{\circ}$ over a horizontal and vertical range of $\pm 20^{\circ}$, with a sampling rate of 60Hz. The subjects sat 63.5 cm from the 48.3 cm screen on which the movie was shown at a resolution of 640x480 pixels.

All gaze data, except for locations which were invalid due to technical or experimental issues, were used in subsequent analysis. That is, the results were generated from gaze points that were not segregated into saccades and fixations. The use of a simple velocity threshold criteria for saccade-fixation segregation (Section 2.5.2) with the cut-off set to 30 degrees sec^{-1} and subsequently labeled saccades removed from study did not change our basic findings, but did improve results across the board for human subjects. Though the effect was small, this finding is consistent with the theory that

visual processing is suppressed during saccades (Burr, Morrone, & Ross, 1994). Since the use of a saccade identification scheme did not impact our results, in this work we omit consideration of saccade identification reasons of economy, with the understanding that the use of an appropriate fixation and saccade identification scheme (given the caveats in Chapter 3) is both relevant and important to a computational model that seeks to describe human gaze patterns.

To assess the performance of human subjects versus chance, it is necessary to define comparative data sets that are basically uncorrelated to human subjects. However, we believe that it is not sufficient to simply sample random points, or to compute statistics over the entire saliency map, to generate our control data. Our set of synthetic data consists of several different types of random gaze strategies:

random filters (rf) – these correspond to a random weight matrix in Equation 6.3-6.4. These are projections that are completely uncorrelated with any events in the visual scene.

random saccades (rs) – these scan paths are created by an algorithm that waits at a given spatial location for some time and intermittently jumps to new locations. The decision to jump is assessed probabilistically, and the distance and angle of jump are generated randomly from uniform distributions.

random physiological (rp) – these scan paths are created algorithmically from physiological gaze measurements using a probabilistic model. The spatial gaze location

of the rp scanpath as a function of the current movie frame number t is $\boldsymbol{g}(t)$, with $\boldsymbol{g}(0)=\boldsymbol{s_0}$ where $\boldsymbol{s_0}$ is the center of the screen. At each new frame the gaze location is updated according to $\boldsymbol{g}(t+1)=\boldsymbol{g}(t)+\Delta(d(t))$, where Δ is a function that takes a step as determined by the distance traveled $d(t)=\|\boldsymbol{g}(t)-\boldsymbol{g}(t-1)\|$. Δ is spatial update in a polar frame, $\Delta=(\,dr\cos(d\alpha), dr\sin(d\alpha))$, where $dr=d(t+1)$, and $d\alpha$ is the change in angle $|\alpha(t+1)-\alpha(t)|$. dr is calculated by a heuristic that samples from the distribution $p(d(t+1)|d(t))$, the dependence of the current velocity on previous velocity. This incorporates the idea that when velocity is high (as in a saccade), it is more likely that movement will continue to be high, and when velocity is low (as in microsaccades during a fixation), it is more likely that movement will continue to be low. The heuristic used is a spill search followed by random sampling: we first locate all time points in the physiological samples where the distance traveled during a given frame was closest to $d(t)$ plus or minus some spill fraction (e.g. 5% of all indices, centered at $d(t)$). We then sample randomly from this collection to get dr. Similarly, the change in angle $d\alpha$ is calculated by sampling from the distribution $p(d\alpha(t+1)|d(t+1),d(t))$, where joint proximity to $d(t+1)$ and $d(t)$ is calculated in the Euclidean sense. The dependence of $d\alpha$ on both previous and current distance reflects the interaction between deflection and velocity in gaze patterns.

The parameters of this synthetic control data are varied in order to span a space of random behavior, and N=5 synthetic sets are generated for each random gaze category. Boundary constraints were enforced so that the range of random position fell within the visible area of the screen and roughly corresponded to human subjects.

6.3.2 Methods

The modified Itti Model with motion, which we employ in our analysis, computes its output on a specific sized image. Since we want our results to be comparable spatially, we first begin by downsampling our input stream so that all saliency maps will match the final size of the Itti model. That is, for raw pixel and Gaussian pyramid techniques, we downsample each image in the stream from 640x480 pixels to 40x30 pixels. This results in a fairly coarse spatial resolution, implying a not inconsequential degree of blurring. However, we have found, with all other parameters held constant, that this blurring increases the performance of our models, likely due to two reasons: (1) it effectively eliminates error due to the inherent inaccuracy of the eye tracking technology used, and (2) downsampling increases the spatial span of our features. The tradeoff between the information lost and spatial range gained is an issue that we hope to address in future work.

Next we apply our various computational models of visual attention to generate features associated with every spatiotemporal point in the visual scene. For every model, the features will be drawn from two time points: {-100ms, -300ms} which straddles the average latency (200 ms) to a visual target by adult observers (Leigh & Zee, 2006). In other words, the features associated with a spatiotemporal point consist of features extracted from the history of that spatiotemporal point, since gaze fixation is not an instantaneous operation, but instead occurs shortly after some salient event or feature is detected. Though we will vary the parameters of our computational models, we will adopt some standards for each model we employ:

raw image patch features – patches are always centered on some pixel, are always square, and contain an odd number of rows and columns. Since our input stream is black-and-white, there is only one dimension associated with each spatiotemporal point: intensity. Each raw image patch is then (*length* x *width* x *point dimensions* x *temporal dimensions*) = NxNx1x2 = $2N^2$ dimensions.

Gaussian pyramid features – we will always employ 3 total levels in our pyramid. If we need to vary the dimensions of this model, we use data from the pyramids in adjacent locations, as we do for raw image patch features. Each pyramid patch is then NxNx3x2 = $6N^2$ dimensions.

Extended Itti Model features – We use modalities of intensity, orientation, and motion. We omit the color modality since the images are black and white. As with the other features, if we need to extend the dimensionality of the Extended Itti Model we use features from spatially adjacent cells. Each feature associated with some spatiotemporal point as computed by this Extended Itti Model is NxNx3x2 = $6N^2$ dimensions.

After we obtain our features, we compute for human and synthetic data sets the optimal filters (as described in Section 6.2.2) for each individual, where an individual is represented by some gaze trajectory over the spatiotemporal scenes. Naturally, we exclude *random filters* from this process. We begin by assembling the set of attended-to features by indexing the features found at each spatiotemporal location that a particular individual's gaze is directed. Next we obtain the set of *not*-attended-to features by

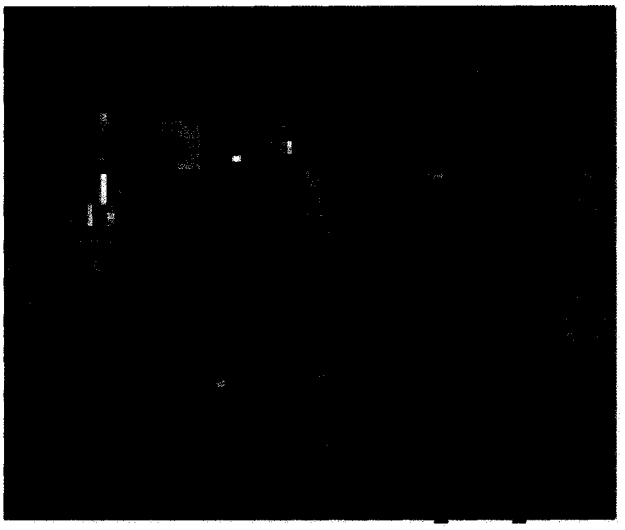


Figure 6.2: Extracting features for attended-to and not attended-to locations. The rightmost box (marked with A) is centered at the attended-to location. The boxes found to the left (unmarked) of the rightmost box are randomly sampled points that are used to generate the *not* attended-to pool.

indexing the features *not* found at the gaze locations of that particular individual (Figure 6.2). We do this by randomly sampling locations that are drawn from the pool of points some minimum distance away from the attended-to location (this distance is, in this work, 31.25% of the number of rows). The minimum distance requirement helps to make the two distributions distinct, as it is known that image patches in natural images are correlated to their distance. 15 *not* attended-to locations are sampled for every attended-to location. Together these sets of features enable us to compute a filter for each individual by finding the optimal projection as given by equation 6.4.

By taking these filters and applying them to the underlying features found at each point in the visual scene, we can obtain saliency maps tuned to each individual as they

watch some particular movie clip. As discussed in Section 6.2.3, we rank order the saliency maps spatially for every time point to obtain rank-ordered saliency maps, and use this as our comparative function. In other words, given an optimal weight W_u computed for some individual u's gaze pattern (see Section 6.2.2), we can calculate the time-varying saliency map $S_u(s,t)$ tuned to that individual u:

$$S_u(s,t) = W_u * F(s,t) \tag{6.8}$$

We then compute, for each frame in the movie, the rank percentile of v's gaze fixation on S_u. That is we find:

$$s_{u,v}(t) = S_u(g_v(t),t) \tag{6.9}$$

$$r(x,thr) = \begin{cases} 0, & x \geq thr \\ 1, & otherwise \end{cases} \tag{6.10}$$

$$R_{u,v}(t) = \frac{\sum_{i \in I} r(S_u(i,t), s_{u,v}(t))}{|I|} \tag{6.11}$$

where $g_v(t)$ is the spatial gaze location fixated upon by user v at time t, I is the set of valid spatial locations in the spatiotemporal scene, and $R_{u,v}(t)$ is the rank percentile score at time t of v's gaze fixation on the saliency map as trained by user u. Since W_u (computed by the Fisher's classification strategy) represents an optimal way of combining features so that locations fixated upon are separated from locations not fixated upon, W_u can be seen as the "gaze strategy" of an individual, and how well separated a set of features are,

then, can be seen as the saliency of those features. The rank ordering technique then becomes a normalization step on top of the saliency calculation to allow different time points to be compared against one another.

Since we are interested in comparing overall performance and group effects, we then generate a receiver operator characteristic (ROC) curve (Zweig & Campbell, 1993) for $R_{u,v}(t)$ as a function of response percentile *rp*. That is, in order to compute the overall goodness of fit of *u*'s model on *v*'s data, we sort all the frames corresponding to $R_{u,v}(t)$ and, in order to make the comparisons more consistent, sample the sorted list at various response percentiles *rp*.

Finally, this information is aggregated into groups and compared. We are interested in both individual effects as well as group effects, and can obtain these measures by utilizing the filters of one individual on other individuals (i.e. *u* and *v* in the above formulation do not have to be the same). For instance, we examine the performance of a human individual's filter on the individual himself ("*self*" datasets), as well as the performance of a human individual's filter on other individuals ("*other*" datasets). We also examine the filters of the synthetically generated random filters, random saccades, and random physiological simulations, when applied to human individuals. We train on the even-numbered frames of one particular movie. This allows us to test upon the odd-numbered frames as well as on a separate movie that does not overlap temporally and which contains different scene content.

6.3.3 Results

By tuning our models to each individual human subject as well as all synthetic data, we are able to generate ROC curves as shown in Figure 6.3 and Figure 6.4. These ROC curves represent the tradeoff between sensitivity (gaze saliency rank percentile) and specificity (user frames percentile). For example, if the model of an individual A were applied to the gaze patterns of an individual B, then if the user frames percentile were 10% and the corresponding gaze saliency rank percentile were 80%, we could say that A's model fits B's gaze patterns quite well because 100-10%=90% of A's data is in the top 20% of saliencies generated for the scene. However, if user frames percentile and gaze saliency rank percentile were roughly equal, we would have an ROC that looks like a straight line, and this would not deviate extensively from what would be expected by random chance.

For our studies we compute cross statistics over different training/testing pairs (i.e. training on one movie, testing on another) and different groupings of human and synthetic data. To study the representational capability of the different feature spaces, we also vary the number of dimensions used for each feature space. This is accomplished by pulling in more information spatially, such that features as not obtained just from a single spatial location, but from a localized spatial neighborhood. The results of these comparisons are summarized in Table 6.1 and reveal several findings.

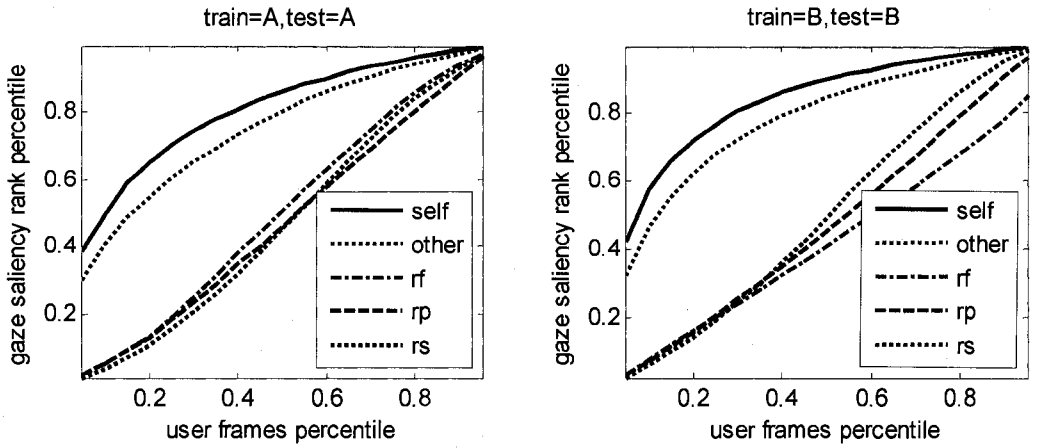


Figure 6.3: Extended Itti Model ROC curves for models trained on gaze patterns from even frames of a movie (A or B) and tested on gaze patterns from odd frames of the same movie. Features are orientation, intensity, and motion over a 3x3 patch across 2 time points. Legend: *self* = models trained on one individual, tested on same individual; *other* = models trained on one individual, tested on all other individuals; *rf*, *rs*, *rp* = models trained on noise, synthetic saccades, synthetic physiological simulations, respectively, and tested on all human subjects.

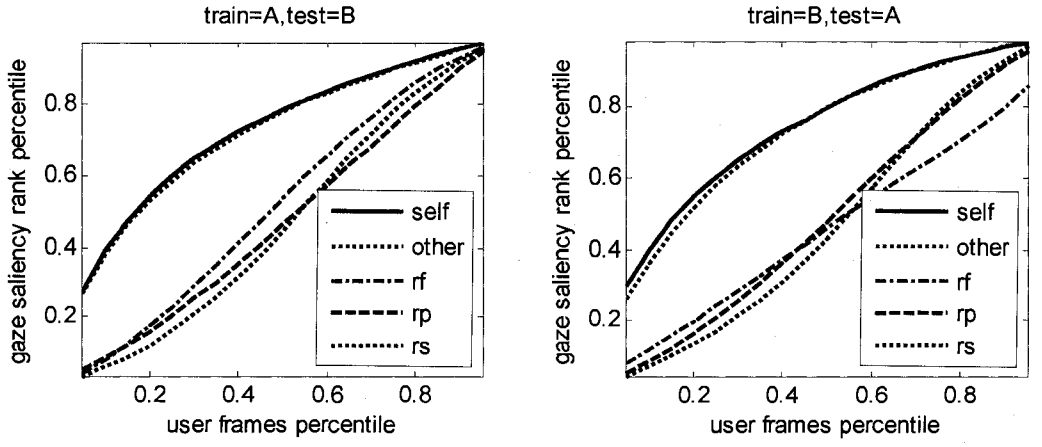


Figure 6.4: Extended Itti Model ROC curves for individual trained on one movie and tested on another movie. Legend as in Figure 6.3.

A on A	L	ND	*self*	*other*	*rf*	*rs*	*rp*
Raw	1	2	74 ± 5	74 ± 5	66 ± 11	50 ± 20	54 ± 23
Raw	5	50	86 ± 3	80 ± 4	53 ± 15	50 ± 10	51 ± 18
Pyramid	1	6	82 ± 3	81 ± 3	47 ± 24	52 ± 19	53 ± 23
Pyramid	3	54	87 ± 2	80 ± 4	52 ± 23	52 ± 9	52 ± 17
Itti	1	6	79 ± 4	79 ± 3	45 ± 24	63 ± 14	58 ± 25
Itti	3	54	86 ± 2	80 ± 4	56 ± 16	44 ± 10	46 ± 14

A on B	L	ND	*self*	*other*	*rf*	*rs*	*rp*
Raw	1	2	77 ± 6	77 ± 6	69 ± 12	51 ± 24	55 ± 27
Raw	5	50	86 ± 3	85 ± 4	55 ± 17	53 ± 9	46 ± 22
Pyramid	1	6	87 ± 3	87 ± 4	47 ± 27	56 ± 22	48 ± 29
Pyramid	3	54	86 ± 3	85 ± 4	48 ± 24	55 ± 9	45 ± 21
Itti	1	6	78 ± 4	77 ± 3	48 ± 22	60 ± 18	58 ± 23
Itti	3	54	78 ± 4	78 ± 5	55 ± 15	43 ± 13	46 ± 17

B on A	L	ND	*self*	*other*	*rf*	*rs*	*rp*
Raw	1	2	74 ± 5	74 ± 5	45 ± 24	45 ± 22	55 ± 23
Raw	5	50	81 ± 3	80 ± 3	47 ± 17	49 ± 7	55 ± 8
Pyramid	1	6	80 ± 4	80 ± 4	46 ± 26	44 ± 19	67 ± 15
Pyramid	3	54	80 ± 3	80 ± 3	62 ± 18	47 ± 9	54 ± 9
Itti	1	6	80 ± 4	80 ± 3	66 ± 11	44 ± 17	51 ± 19
Itti	3	54	80 ± 4	79 ± 4	58 ± 22	43 ± 8	48 ± 13

B on B	L	ND	*self*	*other*	*rf*	*rs*	*rp*
Raw	1	2	77 ± 6	77 ± 6	45 ± 28	44 ± 26	56 ± 27
Raw	5	50	91 ± 2	87 ± 3	51 ± 21	51 ± 11	51 ± 13
Pyramid	1	6	88 ± 3	87 ± 3	45 ± 31	43 ± 25	67 ± 21
Pyramid	3	54	91 ± 2	88 ± 4	65 ± 23	47 ± 9	49 ± 16
Itti	1	6	81 ± 5	79 ± 5	65 ± 11	47 ± 17	52 ± 18
Itti	3	54	89 ± 2	84 ± 4	54 ± 25	49 ± 11	45 ± 13

Table 6.1: Median gaze saliency rank percentiles for variations of computational models of visual attention. Each table represents a single testing/training pairing (e.g. A on B are models trained on movie A and tested on movie B). L is the spatial length (in pixels), and ND is the number of dimensions associated with that particular model's features. Categories are the same as those used in Figure 6.3, and values mean percentages with standard deviations.

First, human subject tuning is better than random even for the largest reported synthetic result ($p<0.05$). In other words, chance, or some general artifact of our processing technique, cannot account for the performance of any model of visual attention that is tuned to human subjects. Second, if we examine the *self* versus *other* human performance across models, we see that, for models trained and tested on the same movie clip, differences appear only as the number of dimensions of the models increase (e.g. differences evidence in Figure 6.3 would not be as pronounced if we were using only 6 dimensions). This suggests that our computational models of visual attention are being tuned to general, rather than specific, strategies at low dimensions. For instance, if we look at the data with the lowest number of dimensions in Table 6.1, that of raw patches of length 1, we can see that *self* performance is equivalent to *other* performance for all cases. This implies that, in this case, tuning the model to a particular individual does not provide greater specificity. When we boost the dimensionality of our features to around 50, however, we see that, for models tuned to particular individuals and tested within the same data set, greater specialization is achieved.

This brings us to the third point: when we apply tuned models to gaze trajectories obtained over different data sets, all differences between *self* and *other* comparisons disappear (Figure 6.4). This suggests that tuning is specific to the spatiotemporal scene over which the model is trained, and that the effects of tuning, when they are apparent, disappear as we move further from the training source (see Figure 6.5). At some basic level, this implies that the actual parameters of these computational models of visual attention are time-varying, suggesting that top-down or contextual effects upon visual attention are observable and significant. In Figure 6.6 we can see this more clearly.

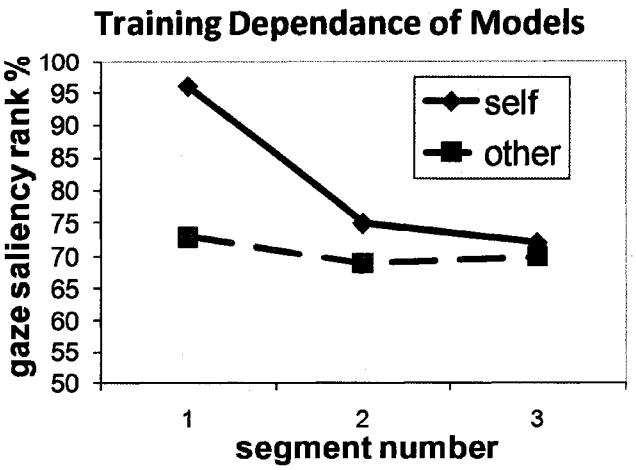


Figure 6.5: Change in model performance as function of distance from the training scene. This model, a raw patch model of high dimensionality (L=11), is trained on subsets of scene A. We take A and divide it into three segments and train on segment 1. The highest matched performance occurs in segment 1, as expected. We note that as we move our testing segment away from segment 1, matched performance decreases with little change to unmatched performance.

When the focus of a human individual shifts from the person who is talking to the person who is being talked to, the model can not readily adapt. In some sense, knowing who is being talked to represents a complex social phenomenon: a truly high-level top-down effect. Our framework thus provides for mechanisms where the weaknesses in a particular visual attention model can be pinpointed and investigated.

Finally, we note that, within our framework, the more complicated Extended Itti Model does not necessarily perform any better, after tuning, than much simpler feature extraction methods. In some ways, this is not unexpected, since biologically-inspired models are not necessarily models that seek to replicate human gaze patterns, but rather are often intended to provide some didactic or theoretical role. Still, it is surprising how

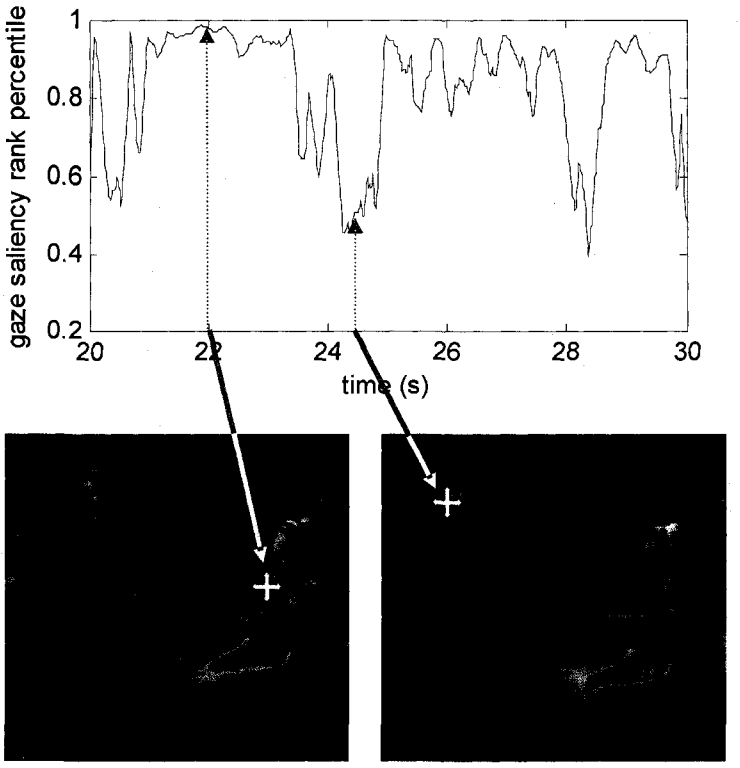


Figure 6.6. Effects of context on model fitting and performance. The top graph represents the time varying gaze saliency rank percentile computed for a human subject (under the 3x3x3x2 Itti Model) applied to his own trajectory. The arrows point to the actual visual scene shown at the associated point in time. The crosses represent the locations where the human subject was actually looking at those times. Note that the model is high-scoring at first, implying that it is well matched to the situation where the blonde rightmost female character is speaking. When the focus of attention of the human subject shifts to the left female character, the model is unable to account for this change.

well a simple method, such as a set of Gaussian pyramid features, can perform even at low dimensionalities (Table 6.1, A on B, 4th row).

6.4 Implications and Limitations

Our system addresses the problem of how predictive computational models of visual attention can be compared with human subjects, and thereby be compared with one another. Validation against human subjects is obviously not the only measure by which computational models of attention may be judged. Draper and Lionelle (2003), for example, evaluate the Itti Model in terms of its sensitivity to similarity transforms. Though Draper and Lionelle frame their investigation in terms of appearance-based recognition systems, their work is applicable more generally. The possibility that known statistical and theoretical properties of the human visual attention system be used to directly evaluate computational models is both intriguing and promising.

The use of random models as controls is one way that such properties could be investigated. The random models used in this current study all share one common aspect: they are computed without regard to absolute spatial and temporal information. Different choices of models which incorporate more information could help determine how particular aspects of the scene interact with the chosen features. For instance, we could randomly choose spatial locations from the set of gaze positions reported in human observers. Such a model would be spatially correlated but temporally uncoupled. Its use as a control would give an indication of the feature dependence on spatial versus temporal information. We could also use human subjects, perhaps engaged in specific tasks, such as target search, as a comparison against the free-viewing experiments we have seen here. Such search-based task patterns would be completely physiological, but the scanning patterns would represent a different underlying motivation.

We should also note that though our formulation is based upon probabilistic intuitions and its application is for the evaluation of predictive models of visual attention, it does not serve necessarily as a generative model for visual attention. In other words, our computational framework is capable of revealing insights regarding how well a model is performing, but it makes no statement regarding what gaze policy should be applied.

An issue that makes it difficult to step directly to some generative model for gaze trajectories in our framework is the fact that visual attention is not stateless. Viewing visual attention as a purely feature-based probabilistic problem leads to behavior that is non-physiological. As seen in Figure 6.7, human eye movements seem to exhibit a great deal of regularity. If we sample from an approximation to the underlying probability distribution, we ignore the strong temporal and spatial correlations inherent to human eye trajectories. It is likely that this framework could benefit from some type of state, as would be found in a Markov model, or in conjunction with the distributions discussed in Chapter 3.

As we have seen, another problem that complicates our analysis is the presence of context-dependent behavior. It is likely that an observer viewing some scene is constantly changing his preferences and objectives, as dictated not only by the scene, but also by some internal mental state. These shifting priorities and desires are likely one factor that contributes to the degradation of our saliency computation as the tested scene becomes temporally further removed. An alternate interpretation of the same effect is one of overfitting. However, if this were the case, the true unmatched normal to normal comparison would be better than what we have reported. As it stands, the results are

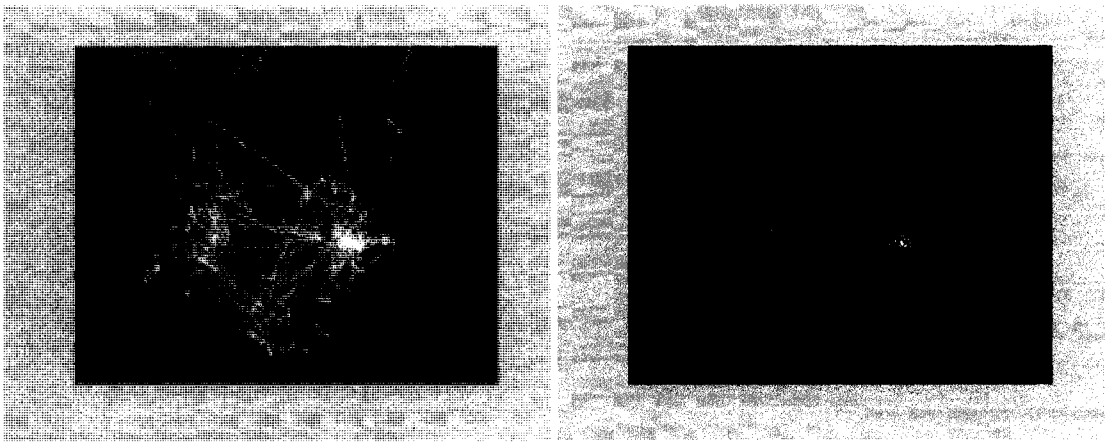


Figure 6.7. Human gaze data (left) and a trajectory drawn probabilistically from an approximation to the underlying density (right). Note the microstructures present in "fixations" of the left image are not present in the right image.

already shown to be significantly different from scene-uncorrelated random models. We should note, however, that the issue of context and top-down effects somewhat hinges on definition. Context effects, in a computational sense, are those effects not adequately represented by the features of a given model. As these features come to be incorporated into a model, their validity as well as the extent of their applicability increases; correspondingly, the rapidity of performance loss due to shifts away from the training source decreases. A model that seeks to represent the scanning pattern of an observer examining a pair of faces laid side by side might have abysmal performance until it incorporates the fact that one of the faces is the mother of the observer.

Our lack of attention to local trajectory statistics and internal mental state is also reflected in our decision to omit the inhibition-of-return mechanism from the Itti Model, possible making its comparison an unfair one. However, it is not clear how inhibition-of-return could be adapted to a dynamic environment with motion, however, since the

addition of a temporal component might suggest that the areas corresponding to inhibited behavior should be time-varying. In addition, a question arises as to how the inhibition-of-return mechanism should be initialized, as the gaze trajectory predicted by the model would interfere with future saliency calculations. This added complexity is likely to be partially why inhibition of return is omitted from many recent investigations of computational saliency (Carmi & Itti, 2006; Itti, 2006, 2005). However, we must admit that use of inhibition of return in the Itti model, which provides some local state and memory, could impact our results, though it is not clear whether it would make the Itti model perform better or worse, or whether other feature extraction methods would benefit from a similar mechanism.

Our custom implementation likely differs in some ways from the implementation available from (Itti, 2008). Because the specifics of the Itti model are clearly defined, with the possible exception of motion, we found it more expedient to implement the model directly. This resulted in a large improvement in the ability of the Itti model to adequately describe the gaze patterns of human observers. The Itti model has evolved substantially from its inception, and it is likely that recent incorporations of signal rectification and local enhancement which, visually, give a more interpretable picture of the salience associated with a given modality, also lead to some loss of information that is not recoverable, and thereby not available for optimization at our classification stage. We have examined the Itti model from multiple angles, under multiple testing conditions, and our results are similar in all permutations.

We should note that we have chosen one particular path in our framework for reasons of computational expediency and illustrative use, but many options exist. In

particular, we have used the notion of saliency as an intermediary step in calculation mainly due to its intuitive nature. However, we are, in fact, evaluating trajectory generators simply by dimensionality reduction over human trajectories – a notion that does not actually require either a true probabilistic underpinning or an explicit formulation of saliency in the manner of Koch and Ullman (1985). There exists an equivalence class of possible saliency schema, the nature, limitations, and capabilities of which we hope to investigate in the future.

We have presented a general technique for evaluating whether a computational model for visual attention behaves in a human-like manner by direct comparison with human subjects. We have shown that distance metrics in image space are insufficient for a general concept of proximity for visual attention, and have developed a classification strategy employing dimensionality-reduction that instead operates in feature space. This classification strategy not only provides a more standardized basis for a notion of salience, but also provides a common interface upon which different models of visual attention may be compared. We have taken a probabilistic version of this classification strategy and transformed it into a dimensionality reduction problem, opening up a broad area of possible inquiry.

By employing our framework, we have shown that the popular, biologically-inspired bottom-up Itti Model, though it serves as a cornerstone for many practical implementations for visual attention, does not necessarily provide any advantage in terms of emulating human behavior in the predictive sense.

In conclusion, we have demonstrated how computational models of visual attention can be developed, applied, optimized, and evaluated. In the next chapter we

will see how the methodology presented in this chapter are not confined to the comparison of computational models, but can also be generalized in order to compare different subjects as well as different groups of subjects.

6.5 Chapter Summary

- We have developed a method for comparing predictive models of visual attention by basing the comparison on the gaze patterns of human subjects.
- We have discussed the problems with simple measures of differences between gaze patterns and have presented a classification strategy which circumvents these problems by grounding the comparison in scene features. This classification strategy employs:
    - Fisher's linear discriminant for separating attended-to locations from locations not attended-to.
    - Rank ordering as a normalization step to ensure comparability across scene frames, different scenes, and different models and individuals.
- We have tested several models of visual attention and have shown that the Itti model, when compared against other simple feature modules, does not show an advantage in the predictive sense.

Chapter 7

Comparing Populations with Predictive Models

Both the pure probabilistic formulation using Bayesian inference and the dimensionality reduction strategy employing Fisher's linear discriminant explored in Chapter 6 are natural methods for tuning computational models of visual attention to the gaze patterns of an individual. Once the model is tuned, the corresponding maps of salience at every point in time and space for that individual are easily generated. We can obtain a measure of how well the model fits by examining the salience at locations where the individual actually looks in comparison to the salience of the locations that the individual does not look.

Once we have a tuned model, however, we are not limited to model-individual comparisons. We can also take this same model and apply it to *other* individuals. That is, we can evaluate how well a particular model, tuned to one particular individual, explains the gaze patterns of other individuals. This process was implicit in the comparison of different gaze patterns against one another in the previous chapter. For example, in the comparisons between randomly generated trajectories and human trajectories, we found that under no underlying model did random trajectories seem to deviate far from chance (Figure 6.3, Table 6.1). We also compared human subjects against themselves and the set of all other subjects, finding that there was an increase in performance when subjects were simultaneously trained and tested on their own scan trajectories. This exploration suggests a methodology whereby the models tuned for one individual are used to gauge the distance between that individual and others.

Furthermore, the results of model cross-application can be aggregated in order to investigate population specific trends.

In this chapter, as a test of our framework and comparative techniques, we apply our methods to the analysis of a population of individuals with autism and matched controls. We know that differences in gaze patterns exist between these two groups both qualitatively (Figure 1.1) and as a result of the high-level analysis conducted by Klin et al. (2002a) which showed that individuals with autism, in comparison to controls, focused more on mouths and objects than on eyes. In this work, we are primarily interested in the implications of cross-population and inter-population statistics upon the developmental and cognitive deficits inherent in autism. This work is based on (Shic, Jones, Klin, & Scassellati, 2006).

7.1 Subjects and Data

The data and subjects in this study were drawn from a subset of the data obtained in Klin et al. (2002a). In this experiment, adolescents and young adults diagnosed with autism (N=10) were matched with a control group (N=10) on the basis of age and verbal-IQ. Other details were the same as that of Section 6.3.1.

As a control against computational bias, several synthetic gaze trajectories were again incorporated into the experiment. These gaze trajectories were uncorrelated with the visual scene and included (1) random filters, (2) random saccades, and (3) random walks (Section 6.3.1).

7.2 Computational Model

Feature Extraction – The features used in this experiment consisted of a linearization of raw patch features drawn from points in history. That is, points of eye fixation corresponding to attended-to locations (and 15 randomly selected points at least 2.9° distant from the actual gaze point for not-attended-to locations) were considered the center of a square area which was further subdivided spatially into a uniform grid of sub-blocks. Each sub-block within the grid was taken to be representative of the underlying spatial content by averaging (i.e. the sub-block represented the corresponding region by a single average intensity), and the set of all sub-blocks associated with selected points in time prior to the fixation constituted the features associated with an attended-to location. The entire grid spanned approximately 9.3° and was divided into 11x11 sub-blocks, sampled at 100ms and 300ms in the past. Temporal sampling was necessary to allow for motion encoding, as the scene was time-varying. Though this feature set was not completely physiological, being coarser in sampling and larger in extent than the fovea, its simple expression struck a useful tradeoff between spatiotemporal extent and computational expedience. Several other feature sets were also tested, including both the multiscale representation as well as the Itti Model (Section 5.2). Neither the use of these other feature sets, nor the variation of their associated parameters within a wide range, impacted the nature of our final results.

Attention Model – Saliency maps were generated by using the method of dimensionality reduction via projection of features upon Fisher's linear Discriminant (Section 6.2.2). Training of models occurred over odd frames of one particular clip, allowing for testing

over the highly-correlated even frames of the same clip, as well as an independent comparison on a completely different clip.

7.3 Comparative Method

Our computational framework provides a method for determining, for some particular individual, the saliency of every spatiotemporal point in the visual scene (Section 6.2.2). If we thus generate a model for an individual A, we can see how well our techniques work by examining the reported saliencies at the locations of A's gaze (Figure 6.6). If our techniques are good, the average saliency at the locations where A fixates should be high. Furthermore, we can take A's model and look at the locations where another individual, B, looks. This gives us a measure of how well the model of A describes the gaze trajectories of B, leading to a natural measure for the distance between the two individuals.

In order to maintain consistency and comparability across all frames in the movies and all individuals we first normalized the saliency values in each frame to a rank percentile (Section 6.2.3). Next, the gaze patterns of a particular individual were indexed into the salience map generated by another individual. From this we were able to obtain time-varying salience records (Figure 6.6). Finally, in order to obtain an overall score representing how well a model matched an individual, the median salience value from the time-varying salience record was taken as representative.

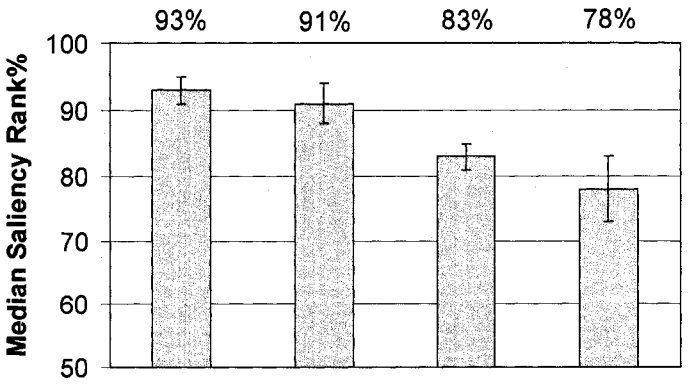
7.4 Results

By applying models tuned for each trajectory (both human and synthetic operating over two movie clips) to every other trajectory in our data set, we were able to obtain a large number of cross-trajectory comparisons. By aggregating the data into groups we obtained the statistics of Figure 7.1 & Figure 7.2.

The application of our framework leads to several results. First, all applications of a human's model to a human's gaze trajectory lead to performance much better than those obtained by random chance (52±13%, N=600), as developed by synthetic gaze trajectories (Figure 7.1 & Figure 7.2; $p<0.01$). This suggests that both individuals with autism and control individuals rely on some common scanning approach, implying the existence some core human strategy. Furthermore, this result suggests that it is unlikely that a methodological bias exists in either the learning technique or the feature representation. When a model is trained on one movie and applied to another movie, we get a drop in performance.

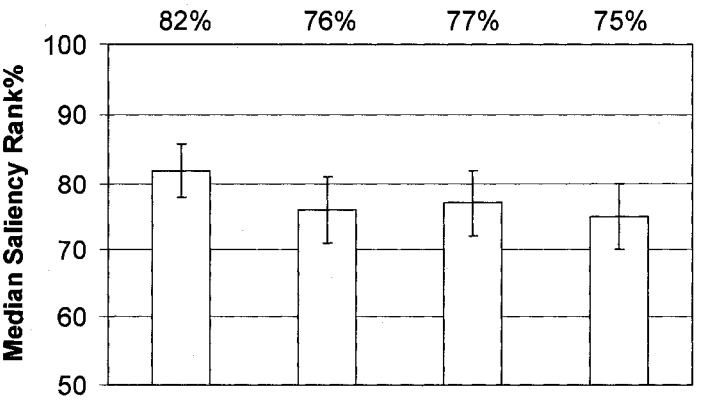
Second, the extremely high matched-application (control on self and autism on self groupings) within-movie scores (Figure 7.1) suggest that each subject relies upon some specific individual strategy. This specific individual strategy does not seem to transfer across scenes, as demonstrated by matched comparison score drops as we move from within-movie comparisons to across-movie comparisons, suggesting that top-down or contextual influences on gaze strategy are significant.

Third, as highlighted by Figure 7.2, control individuals, who are taken to be socially more typical than individuals with autism, exhibit much greater coherence ($p<0.01$) in terms of attraction to underlying features than cross-population cases that



Model of	Control	Autism	Control	Autism
Gaze of	Self	Self	Self	Self
Movie	Same	Same	Diff	Diff

Figure 7.1: Self-tuning comparisons across movies. Results (N=10 each condition) for models trained on one individual (control or autism) and tested on the gaze patterns of the same individual (watching the same movie or a different movie). Error bars in standard deviations.



Model of	Control	Control	Autism	Autism
Gaze of	Other Control	Autism	Control	Other Autism
N	90	100	100	90

Figure 7.2: Cross-tuning comparisons within the same movie clip. Models for the gaze of controls describe the gaze of other controls better than the any cross-population comparison that involves autism, including autism models applied to the gaze of other individuals with autism.

involve individuals with autism. This suggests that the strategies of controls transfer well to other controls, but that the strategies of individuals with autism do not transfer to the same degree to either normal individuals or even other individuals with autism.

7.5 Implications and Limitations

The original Klin et al. (2002a) study found that individuals with autism spent more time focusing on mouths, bodies, and objects, whereas controls spent significantly more time looking at eyes. In terms of elementary features, eyes vary the least; objects vary the most. Thus our results in this chapter could derive specifically from this disparity. If eyes vary the least, and controls focus on eyes much more often than individuals with autism (the difference between eye fixation time fractions between the two populations exceeds 40%), we would expect a higher correspondence among control individuals. Similarly, if features associated with bodies and objects vary most, we would expect individuals with autism to exhibit fine tuned strategies specific to particular objects or image characteristics not generally found elsewhere. If these strategies are extremely fine tuned, they cannot transfer to other individuals.

The disadvantage of a predictive model analysis, compared to, for example, ROI based analysis (Chapter 4) or descriptive model analysis, is that much of the internal circuitry after optimization is impenetrable. For instance, we can frame our results in terms of semantic labels associated with subject fixation. However, the converse, predicting high level implications from low level aggregate effects, could prove very difficult. On the other hand, since we do have as many time-varying salience records as

we have comparisons, it is possible that by pinpointing locations of mutually high salience we could discover classes of highly correlated specific gaze behavior. The use of our comparative techniques in this manner is a future avenue to be explored.

The advantage of featural level analysis is that preexisting labels with associated semantic implications are not assumed. If the underlying featural representation associated with a particular computational model of visual attention is sufficient to represent some common underlying strategy within a population, our techniques should uncover this fact. In this investigation we have uncovered two tiers of shared strategies. The first tier represents the underlying gaze patterns associated with the scanning behavior of all humans, mechanisms possibly hardwired into the early visual system. The second tier is found between controls, possibly representing typical development versus early derailment as predicted by enactive mind theory (Klin et al., 2003). Finally, the ability for models to match specific individual preferences suggests that order *does* exist in the gaze patterns of individuals with autism, suggesting that if early derailment of social skill development is occurring, it is replaced by some other set of visual behavior that likely reflects a unique cascading specialization.

There are several additional possibilities that could be accomplished with this technique. First, since this comparative strategy provides a means by which one individual's gaze model can be tested against multiple individuals, it may be possible to compute a distance that is not formulated in a group comparison, but rather at the level of an individual to individual. We could imagine taking these distances and using them to see if they match with, for example, cognitive or social measures of functioning. A simple exploration would be to examine the level of impairment or the degree of autistic

severity in the individuals with autism as a function of their average model explanatory power for typical individuals. Another possibility is that clusters, given the matrix of differences between individuals, could be derived automatically. This would offer the possibility of subtyping individuals, giving us a means by which treatment could be tailored or endophenotypes for genetic analysis could be uncovered.

7.6 Chapter Summary

- We have adapted the comparison techniques used for comparing models against one another to compare individuals to other individuals, obtaining a measure that gives us a meaningful notion of the distance between two gaze patterns on a scene.
- We have shown that models trained on human subjects are more closely matched with other human individuals than those models trained on synthetic gaze data. This suggests a core human strategy for scanning on the presented scenes.
- We have shown that models trained and tested on the same person perform much better than models trained on one person and tested on another person, suggesting specific strategies for each individual.
- We shown that individuals without autistic psychopathology exhibit much greater coherence in terms of their attention to features than those individuals with autism. Results suggest that the gaze patterns of individuals with autism are as far removed from each other as they are from those without autism.

Chapter 8

Descriptive Computational Models of Visual Attention

Descriptive computational models for visual attention are designed to provide maximal insight as to the underlying process occurring at the point of regard. A common technique is to extract features of the scene in which we are interested and to match those features to the locations where individuals look. This type of analysis is best suited for low-level visual properties which reflect spatiotemporally localized image statistics, such as the amount of color, motion, contrast, spatial frequency, or orientation power spectrum at locations in the observed scene. Thus, region-based and descriptive computational modeling can be viewed as two separate avenues for approaching the question "what are subjects looking at?"

From one side, region-based modeling provides a method for experimenters to test high-level theories regarding visual scanpaths, allowing researchers to leverage clinical, cognitive, and psychological insight. From the other side, descriptive computational modeling provides a method by which the psychophysics of the locations scanned can be studied, allowing the statistical properties of the scenes viewed by subjects to come to fore. The two methods are by no means mutually exclusive. It is certainly the case that one could go to a stimulus to be presented and draw regions around all the areas that appeared highly unique from a perceptual standpoint. Similarly, one could create a computational model that finds objects or faces and uses the areas found in a region-based analysis. However, in their traditional roles, region-based analysis and

descriptive computational modeling can provide complementary pieces of information, both with the end goal of describing the underlying motivation which drives an individual to look at one location over another.

There are some tradeoffs to consider when employing a descriptive computational model of visual attention. First, if the goal is to investigate certain perceptual abnormalities, we could construct sensory experiments aimed at testing those specific psychophysical effects. This would lead to results that are less ambiguous and often more interpretable. However, this approach has the drawback of not testing perception in the natural environment, making it difficult to generalize the role of, say, deficits in a specific task to everyday functioning. If we want to access the usage of basic perceptual features in an ecologically valid way (i.e. in situations closely aligned with real social experience), we need mechanisms by which scenes viewed by individuals in general can be decomposed into elementary properties. That is, we need a low-level analogue of high-level interpretations of abnormal gaze patterns. To do this we can follow the route of Neumann, Spezio, Piven, & Adolphs (2006) and Parkhurst and Niebur (2003): we can employ descriptive models of visual attention in the analysis of scene content.

Evaluating gaze patterns in terms of elementary features can provide measures for comparing the preferences for low-level modalities in one group against another. However, this only provides a perceptual baseline for a set of cognitive processes affected by multiple aspects. To gain access to higher-level aspects we must take into account context, where context is here operationally defined as those factors not accounted for by the particular computational framework. Manipulating the context of a

scene gives us a direct quantitative measure, in terms of effects on basic perceptual properties, of that contextual factor.

In this chapter we will examine the gaze preferences of children with and without autism spectrum disorder (ASD). We will manipulate two contextual effects known to impact visual attentional response: scene orientation (e.g. face inversion) and sound (e.g. loud noises causing alarm). We will gauge the contribution and impact of the contextual modification of scene orientation and sound. We will show results that are consistent with previous results found in literature and which also provide interesting avenues for future exploration. This chapter is based on work published in (Shic et al., 2007).

8.1 Evaluation Metrics

An individual will look at particular points in space over time; we are interested in matching these points to their associated low-level perceptual interpretations. In order to accomplish this, we employ the internal representations of the Extended Itti Model (Chapter 5.2) which decomposes the scene into elements of intensity (contrast), orientation, color, and motion. We begin with a conspicuity map, here generically defined as $\overline{V}(s,t)$ for some feature V, spatial location s, and time t. In order to obtain comparability for all time points and all modalities, we first normalize the values of the conspicuity map by rank ordering all the spatial values for a given time, to obtain the rank-ordered conspicuity map $\overline{V}_r(s,t)$:

$$r(x,thr) = \begin{cases} 0, & x \geq thr \\ 1, & otherwise \end{cases} \tag{8.1}$$

$$\overline{V}_r(s,t) = \frac{\sum_{s' \in S'} r(\overline{V}(s',t), \overline{V}(s,t))}{|S'|} \tag{8.2}$$

Note that though this is a similar process to the rank ordering used in Chapter 6.3.2, when rank ordering is applied to the obtain smoother evaluative statistics for prediction, in this case we are rank ordering the constituent maps themselves rather than the saliency map. We do this because it is the utilization of these maps that is of prime importance in this application.

Given the gaze patterns of some individual i, $\boldsymbol{g}_i(t)$, we can obtain the perceptual usage of V at time t by i as $v_i(t) = \overline{V}_r(\boldsymbol{g}_i(t),t)$; we can in turn use this time-varying perceptual usage score to compute the aggregate perceptual score $p_{v,i}$ = median$_t(v_i(t))$.

8.2 Subjects and Data

20 typically-developing (TD) children and 44 children diagnosed with autism spectrum disorder (ASD) participated in this study. The age of the TD population was 44.9(5.9); the age of the ASD population was 43.9(8.1) months. The diagnosis of ASD was determined by expert clinicians as part of a comprehensive clinical examination at the Yale Child Study Center.

Each child was accompanied by his parent into the room where the experiment was conducted. Children with sufficient neck support sat in a car child seat strapped to a chair; younger children were held over their parent's shoulder as the parent sat in a chair. A monitor, centered with the eye-line of the child, was mounted 75 cm from the child's face. The child's gaze patterns were tracked using a commercial eyetracker from SensoMotoric Instruments (iView X RED) at 60Hz.

The experiment in this study was embedded in a large run of several different experiments so as to minimize the overall amount of time spent positioning and calibrating the child. The child saw 4 movie clips, with each clip measuring approximately 24x18 (width x height) visual degrees and lasting for 30 seconds. All clips depicted a natural interaction between an adult caregiver and a child (e.g. playing with a toy) (Table 8.1, Figure 8.1). Each clip was shown in one of four conditions representing the modulation of two variables: orientation (inverted or upright), and sound (mute or sound).

During a single experimental session, the clips were always presented in the following order: inverted mute, inverted with sound, upright mute, upright with sound. Each clip presented to the child during a single session contained different scene content. However, a child could engage in multiple sessions, with each session conducted on a different day. In cases where children engaged in multiple sessions, they would see the same scene content on different days, but would never see the same scene-condition pairing twice, as the scene content would be rotated amongst the conditions. Clips were rejected from analysis if they contained less than 10 seconds (600 points) of valid eye-tracking data (ASD 22 clips; TD 8 clips); typically this rejection occurred due to the child

Scene	Description
1	Child tries to put objects into a colorful container; Caregiver instructs child to insert toys in holes by pointing gestures
2	Depicted in Figure 8.1; Child lifts head to speak to caregiver; Child opens container and drops in toys
3	Child offers toy to caregiver; Caregiver teaches child name of toy; Child repeats name of toy while continuing to play
4	Child enters scene with Caregiver and a mechanical toy that spits out colored balls; Child plays with back to camera (obstructing toy)

Table 8.1: Descriptions of the four video scenes shown to children

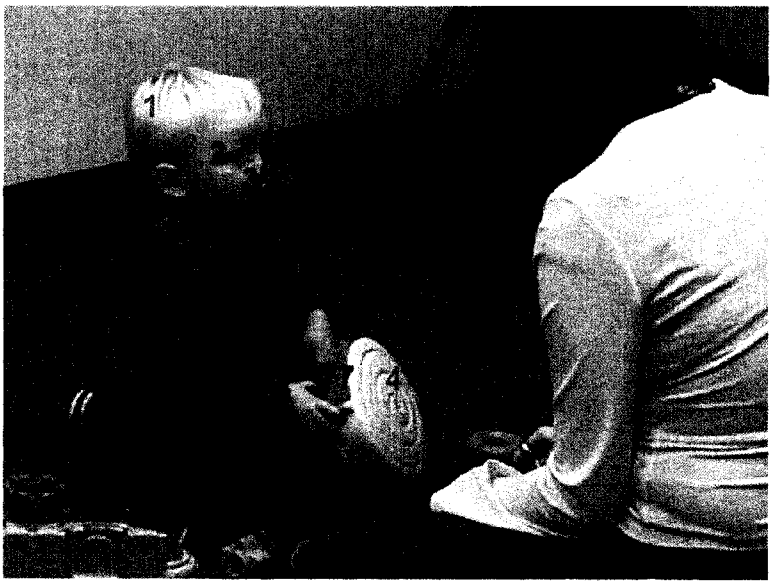


Figure 8.1: Example of one frame from a scene shown to children (in the upright with sound condition), with the gaze locations of ASD individuals (red) and TD individuals (green) for that frame overlayed.

affect or inattention. In total, the ASD population contributed 157 clip viewings over 49 sessions; the TD population contributed 88 clip viewings over 24 sessions.

8.2.1 Data Processing and Analysis

Features were extracted from the movie clips in the manner described in Section 8.1. For each clip viewing, the gaze patterns of the children were mapped to the associated features 200 ms in the past for reasons described previously (Chapter 6.3.2).

Data processing yielded an aggregate perceptual score for each modality (intensity, orientation, color and motion), for every clip viewing. Each modality was analyzed independently using a univariate analysis of variance with factors: diagnosis (ASD or TD), orientation (upright or inverted), sound (mute or with-sound), and the specific scene content (a subset of 4 possible scenes). Age was listed as a covariate, as the range in ages of both ASD and TD populations was large. Because the perceptual score for a particular modality was tightly coupled to the scene (i.e. when analyses were originally conducted, the effect of specific scene was by far the most significant effect), scenes were analyzed together only when the set of scenes together did not register a significant between-subject effect. In the event of multiple choices amongst scene combinations, the combination resulting in the greatest number of subjects was retained. The data used in this study are summarized in Table 8.2. The pattern of aggregate perceptual scores was tightly coupled to the scene content, as shown in Table 8.3.

After controlling for scene content, no significant effects of diagnosis, scene orientation, sound, or age were found for color and orientation. However, significant differences were detected for intensity and motion. For intensity, there was a main effect

Modality	n_{asd}	n_{td}	n_{upr}	n_{inv}	n_{snd}	n_{mute}	n_{scene}
Intensity	122	67	97	92	87	102	3
Orientation	79	45	66	58	49	75	2
Color	43	22	31	34	38	27	1*
Motion	76	42	62	56	60	58	2

Table 8.2: Data Characterization. Variables: Number of clip-viewings: n_{asd} (ASD), n_{td} (TD), n_{upr} (orientation upright), n_{inv} (orientation inverted), n_{snd} (with sound), n_{mute} (no sound); n_{scene} (number of scenes combined for analysis (based on comparable means)); *for color no scenes were comparable to any other.

scene	Intensity			Orientation			Color			Motion		
	μ	σ	N	μ	σ	N	μ	σ	N	μ	σ	N
1	.35	.10	65	~~.67~~	~~.07~~	~~65~~	.70	.11	65	~~.77~~	~~.08~~	~~65~~
2	.35	.07	62	.55	.05	62	~~.78~~	~~.10~~	~~62~~	.70	.06	62
3	~~.49~~	~~.07~~	~~56~~	~~.67~~	~~.10~~	~~56~~	~~.46~~	~~.08~~	~~56~~	.69	.08	56
4	.35	.04	62	.55	.07	62	~~.65~~	~~.06~~	~~62~~	~~.55~~	~~.03~~	~~62~~

Table 8.3 Perceptual Scores of Modalities for Each Scene. Scenes not comparable to other scenes were removed (crossed-out). Reported means are collapsed across conditions and diagnoses. See text in this section for explanation of exclusions.

of diagnosis (ASD vs TD) ($F(1,188) = 5.7$, $p<0.05$) and scene orientation (upright vs inverted) ($F(1,188)=11.6$, $p<0.001$), and an interaction for diagnosis x scene orientation ($F(2,188)=6.8$, $p<0.01$). The effects of sound (with-sound or mute) and age were not significant.

To examine the nature of the interaction, simple between-group comparisons for each of the four conditions (i.e., inverted-mute, inverted-sound, upright-mute, upright-sound) were conducted. The comparisons indicated that ASD and TD groups differed significantly only in the upright-mute ($F(1,41)=11.94$, $p < .001$) and upright-sound ($F(1,49)=12.13$, $p <.001$) conditions, but not the inverted-mute ($p>.28$) or inverted-sound ($p>.77$) conditions. Furthermore, we compared intensity scores within each group in the upright and inverted conditions. These within-group comparisons indicated that toddlers with ASD were not affected by scene inversion ($p>.43$), but in TD toddlers the salience of intensity increased significantly when the scenes were inverted ($F(1,66)=18.55$, $p < .001$).

For motion, there was a main effect of diagnosis ($F(1,117) =6.3$, $p<0.05$), scene orientation ($F(1,)=4.0$, $p<0.05$), and sound ($F(1,117)=9.2$, $p<0.01$). There were no significant interactions between the factors, but the effect of age on the salience of motion was significant ($F(1,117)=5.6$, $p<.05$). Toddlers with ASD were less sensitive to motion cues, regardless of the condition (i.e., scene orientation or presence/absence of sound). All toddlers tended to be more sensitized to motion in the sound than no sound conditions and when the scenes were inverted as compared to the upright. Results are summarized in Figure 8.2.

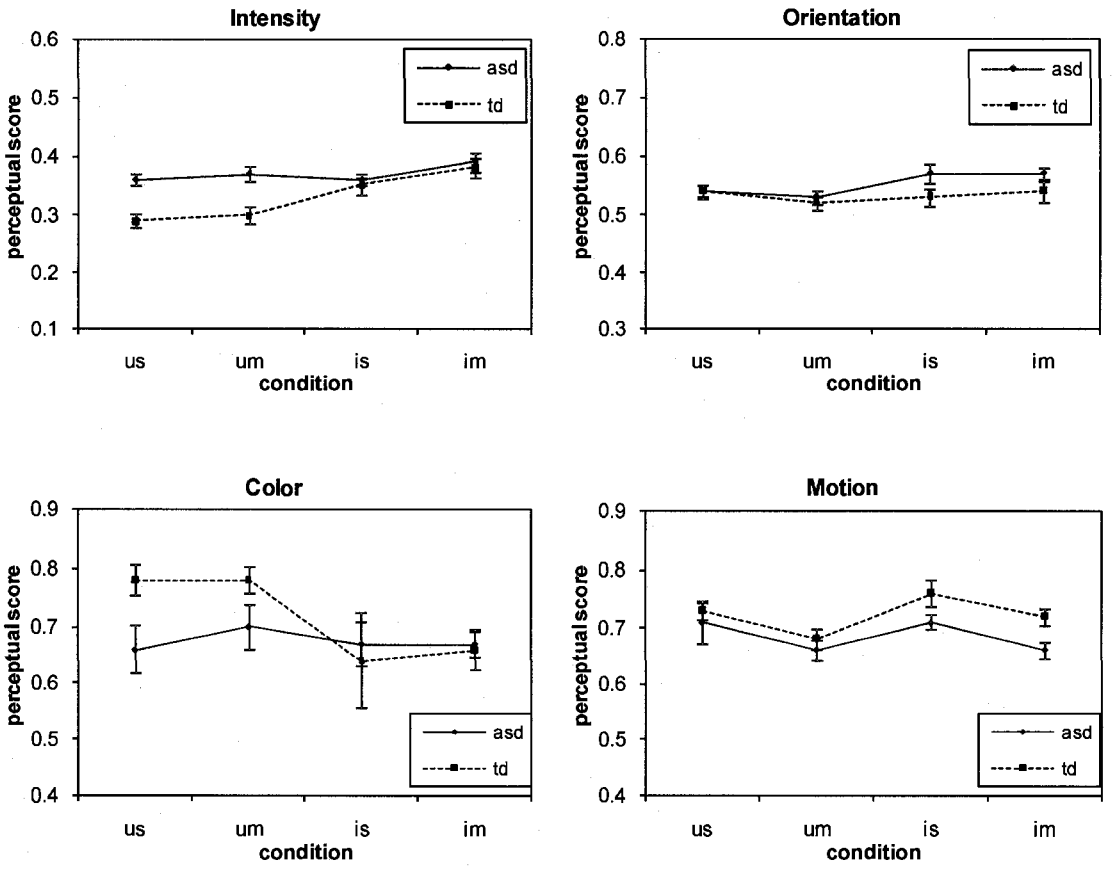


Figure 8.2: Aggregate perceptual scores by diagnosis for each modality. The x-axis is ordered from least disturbed to most disturbed. The categories are: inverted muted (im), inverted with sound (is), upright muted (um), and upright with sound (us). Only the differences for intensity and motion were significant. Note that the scores for each modality are displayed on different scales but over the same range. Bars are ± one standard error.

8.2.2 Implications and Limitations

At the baseline, that is, in the least disturbed state, typical controls use less intensity information than children with autism. Upon scene inversion, however, the utilization of intensity by typical controls increases to the point where the ASD and TD populations are

indistinguishable. By comparison, for children with autism, no effect of scene inversion is observed. This result is evocative of experiments by Tantam et al. (1989) where he found that face inversion decreased face recognition performance for typical individuals, but not individuals with autism. Furthermore, this result is consistent with evidence for local features preference or global processing deficits in autism. If typically developing controls, in the least disturbed condition, are using configural or holistic processing, it is expected that by inverting the scene we would disrupt the use of specialized visual strategies, leading to a commensurate increase in purely bottom-up perceptual usage. By contrast, if individuals with autism have a preference for local visual processing then we would expect that scene inversion would largely leave the attentional strategy of these individuals intact, as is supported by our data. We note that this possibility is not the only interpretation, for it is possible that a disturbed contextual factor could in turn be replaced by some other contextual factor.

For motion, under all conditions, we note that children with autism use more motion information than their typically developing peers. This result is consistent with results from literature demonstrating motion processing deficits in autism (Blake et al., 2003; Frith, 2003b). In addition, the covarying effect of age on the motion usage in typical children is supported by research showing decreasing thresholds for motion detection with increasing age (Braddick, Atkinson, & Wattam-Bell, 2003). Finally, the trend for both typically-developing children and children with autism was for greater motion usage in the presence of sound. In many of the scenes shown, a large amount of motion accompanied large changes in sound. It is possible that the presence of sound increased the urgency of motion.

We also note that the specific scene content under consideration plays a critical role in the reported perceptual usage. In the scenes presented in this experiment, the variations in choice of focus, the different actions that were performed, the various implicit object-people interactions, all could have had dramatic impacts on the expected perceptual values. This dramatic effect of scene content points to the inherent imitations of computational models of visual attention which try to model human gaze explicitly without optimizing for individual biases. This is a lesser problem for the use of computational feature extraction techniques in the *evaluation* of basic perceptual modalities. Nonetheless, the quality of the evaluation of modality usage is only as good as the techniques employed. If our selected computational model could adequately predict attentional salience, we would see not only that context effects contributed only in a minor fashion to perceptual usage, but that the computational model accounts for a majority of an individual's gaze patterns. This is, however, not the case in this study, as the results for perceptual usage based on our computational model are nowhere near the maximum.

The present study is limited in many ways. First and foremost, the complexity of the experimental design necessitates a large sample size, whereas the present study only has a moderate-sized set of subjects. For this reason the results in this chapter should only be taken as preliminary. With a greater sample size it is possible that interesting effects could be uncovered for color and orientation modalities, both of which yielded no significant results. Second, the age range of subjects is extremely wide, necessitating the use of age as a covariate. This was a problem addressed in a follow-up study (Shic et al., 2008) which constrained the age ranges to a much smaller range. The results of this

study are not addressed here, as the results were very similar and are still in progress. Third, the appropriateness and quality of the Itti model feature calculations was not conducted. It is certainly the case that the Itti model, as a model of *visual attention*, does not match up with human gaze patterns, though the specific saliency values are better than chance (Shic et al., 2006). Without optimizing the parameters of the model (Shic et al., 2006; Shic & Scassellati, 2007), it will necessarily be a poor fit. It is, however, in our opinion, a sufficient construct for serving as an *evaluative* model of visual preference.

We have taken a computational model of visual attention and employed its internal mechanisms as a strategy for evaluating gaze patterns in terms of elementary perceptual features. We have adapted this model with a framework for evaluating context by scene manipulation, and used this framework to evaluate the perceptual strategies of individuals with autism as compared with those of typically developing controls. Through these techniques, we have generated several results, framed in terms of perceptual utilization, which are consistent with other results from literature. We find that children with autism use less motion information than their typically-developing peers, consistent with motion deficits shown in autism. We find that children with autism, in terms of the perceptual utilization of intensity, are more resistant to scene inversion, supporting the role of local visual processing preferences in autism. We also note that motion effects consistent with developmental trends in age and known interactions with sound are shown.

8.3 Chapter Summary

- We have applied a computational model of visual attention (the Itti model) in the evaluative sense, that is, to gain insight into the perceptual basis by which visual attention may be allocated.
- We further employ a technique of context modulation where aspects of the scene are manipulated and the effect of this manipulation determined in terms of attention to perceptual features.
- We have used this model to study the perceptual properties that children with autism attend to as compared to typical children.
- We find some indication that typical children may adjust their attention to perceptual properties upon scene inversion, whereas children with autism seem fairly invariant to scene inversion.
- We find that children with autism look at areas of greater contrast (intensity in the Itti nomenclature) than typical children, and that they look less at areas of motion.

Chapter 9

Summary

We have explored a wide range of computational and analytical methods for eye-tracking analysis. These methods do not stand as separate components of an analytical framework, wholly removed from one another, but rather represent complementary and reinforcing methods and techniques. Region-based analytical approaches offer us the ability to examine information at a high level, asking questions about where a subject is looking and what he is looking at, from the top-down. Descriptive computational methods give us the ability to look in the other direction, examining preferences for elementary perceptual features, from the bottom-up. Predictive computational methods allow us to look at differences between subjects and groups in an automatic fashion, and give us the ability to make comparisons across multiple domains in a fair fashion. Finally, the linear and power-law models that we have used to examine the parameter problem in fixation identification simultaneously question the assumptions held by traditional eye-tracking analysis and help to clarify the spatiotemporal distribution of eye-tracking data.

It is not by chance that so much of our work has practical implications. At the beginning of our research, we operated only over scanpaths given to us by generous collaborators; we did not collect this data. At this point we believed that eye-tracking data was simple and clean, with little ambiguity; but we were wrong. When we began to collect our own data, we realized that we needed to process many things differently. Some changes were big, some were small. As is the case with proprietary software, a

small change is as easy as a big change: in both cases, you can't do it. For this reason we had to reengineer and rewrite the entire suite of eye-tracking analysis and investigative tools from scratch (Chapter 2). Though it took some time, the insights provided by this process have been invaluable. As with most branches of packaged and commercial diagnostic technologies, you don't question basic assumptions until they stop working.

And similarly, it is perhaps by momentum that the nature of our work reflects increasingly clinical and psychophysical exploration. We actually began our work in eye-tracking with the framework for evaluating computational models of visual attention (Chapter 5 and 6). This lead us to investigate how these models could be used in order to determine the distances between the gaze patterns of different subjects, as given by our work with predictive models (Chapter 7). From here, driven by a need to better tackle questions in autism, we moved simultaneously through the use of computational models in the evaluative sense (Chapter 8) and more high-level region based analysis (Chapter 4). And it is while we were trying to pick a best set of parameters for fixation analysis that we discovered methodological flaws our basic assumptions (Chapter 3).

And what have the travels we have undertaken said about autism? Along this same time-line, we first found, though our use of predictive models of attention as a measuring stick of distance, that gaze patterns of individuals with autism were as far from each other as they were from controls. This implies a heterogeneity in the disorder, because, while the controls tended to look at the same places in a scene, individuals with autism did not. Each one of them was in their own space, a result consistent with etiological, behavioral, and cognitive variability results that have lead some researchers to refer to "autisms" rather than "autism" (Geschwind & Levitt, 2007). These individuals

were roughly 16 years of age. When we began to consider these computational models in the evaluative sense, we found some evidence that children with ASD seemed to look at areas of contrast more and motion less. These children were about 4 years old. When we examined the distribution aspects of scanning, as motivated by work in fixation identification, we found that the scanning distribution over blocks and faces by typical toddlers was very different, but, despite the vast differences between blocks and faces for ASD toddlers, the spatial measures scanning between these classes of images was similar. These toddlers were two years old.

And while this work is primarily focused on computational methods, it is perhaps a little difficult not to talk about autism, given that it touches our work at almost every level. From our work in region-based analysis we know that there is a difference between children with ASD at two years and at four years. Attention to and exploration of faces in general is lower in the older ASD children than in the younger ASD toddlers, and these difference are in stark contrast to the opposite effects observed in their typically developing peers. It is quite premature to speak of conclusions to be drawn as the result of the findings in this thesis. However, we can offer the following view.

If a baby enters this world, into this, as William James puts it, "great blooming, buzzing, confusion", without ever having seen a face, how is it that every typical developing child eventually becomes enamored with others? This is a very old question, and the work in this field is as profound as it is prolific (e.g. see Goren et al., 1975a; Haan et al., 2002a; Meltzoff & Moore, 1977; Valenza et al., 1996a). If a child begins with disturbed innate mechanisms for predisposing him towards looking at faces, this could lead to decreased social motivation as the child grows older. As motivation

decreases, looking at relevant aspects of faces decreases. By contrast, perceptual or basic qualities of scenes would increase in prominence. As the typical social motivations become a unifying force for drawing attention in typical individuals, the atypical individuals without this motivation begin to scan what is left over. Since by definition, a unifying marker for attention draws typical individuals to the same location, the lack of attention towards this marker would necessarily imply a greater deviation to other factors of the scene. But then, why does there seem to be less attention towards motion for ASD children in our descriptive computational modeling? One possibility is that the objects that are moving in those scenes are exactly those tied to human actions: the child pushes a button, the caregiver waves her hands. It seems then, this explanation describes how an initial insensitivity towards faces, found at two years, could lead to a lack of attention to the motions of people in videos as well as towards key areas of the face, found at four years. From here, this lack of a cohesive social glue in the viewing of social scenes leads to a greater heterogeneity when comparing those scanning patterns on those scenes, found at 16 years.

Other explanations for the schedule of results we have found are possible, however. For instance, in a typical scene, there are perhaps a few moving objects and many non moving objects. In these situations, then, motion provides a unifying cue for attention—to those whose preferences for motions are similar. It is possible that with an initial insensitivity to biological motion or an atypical preference for physical motion contingencies, one would attend less to looming or speaking faces, triggering the hypothesized pattern of events in the previous paragraph (Blake et al., 2003; Klin, Lin, Gorrindo, & Jones; Lin et al., 2007). Similarly, it is possible that when children with

autism attend to motion they attend for less time, a view consistent with both the perceptual abnormalities and pervasive inattention to faces that we have observed.

In any case, however, it is clear that considering the impact of initial biases on an individual's development, and considering the developmental trajectory as a whole rather than as isolated incidents (Karmiloff-Smith, 2007; Klin et al., 2003), will be an essential part of piecing together the differences we have observed, though our layers of new eye-tracking methods and approaches, into a coherent picture describing the social epigenesis of autism. Our results suggest that the differences between individuals with autism and their typically developing peers are detectable in the distributional aspects of gaze patterns towards social and non-social stimuli by two years of age. From here, a host of low-level differences, such as a greater attention to contrast and lesser attention to motion, as well as high-level differences, such as a pervasive inattention towards faces, emerge in autism by four years. This atypical trajectory eventually ends in a splintering of the typical gaze strategy, with a great heterogeneity found in the strategies used to scan social scenes in adolescents and young adults with autism. The techniques we have developed thus allow us to cut across development, giving us clues as to the biases which may impact the progression of autism, and furthermore provide for us a rich set of perspectives from which to watch this progression unfold.

Our multileveled approach towards analysis can help us isolate the differences between groups—but this is just the initial step. In order to tie these differences to the underlying processes more precisely requires an extension of not only analytical techniques but also experimental design. To better understand how the parameters of our distributional, region-based, and perceptual models relate to the actual recognition of

stimuli, for example, we could employ the visual paired comparison paradigm (Fantz, 1964; Richmond, Colombo, & Hayne, 2007; Richmond, Sowerby, Colombo, & Hayne, 2004). In order to determine whether perceptual abnormalities in natural settings correspond directly to perceptual irregularities at the elementary level, we could create a sequence of tests, using stimuli that span the range from simple contrast and motion gratings to complex and subtle social scenes. And, of course, the association of the measures obtained from our eye-tracking methodologies with standard psychological and psychopathological assessment scores should prove to be invaluable in extending our research methods to a truly diagnostic technology.

The methods that we developed in this work can be seen as a series of techniques and approaches that can be used to dissect eye-tracking data at many different levels, not only for autism, but for many clinical, psychological, and psychophysical applications. In many cases these methods answer questions in a very different way than traditional methods would, and can also be more convenient, more complete, or more flexible. It is our intent to continue developing the methodology proposed here, for the express purposes of deciphering the mysteries of psychopathology, so that we may not have just a window to the soul, but multiple windows, of different shapes, sizes, and manifestations.

Bibliography

Aks, D. J., Zelinsky, G. J., & Sprott, J. C. (2002). Memory Across Eye-Movements: 1/f Dynamic in Visual Search. *Nonlinear Dynamics, Psychology, and Life Sciences*, *6*(1), 1-25. doi: 10.1023/A:1012222601935.

Althoff, R., & Cohen, N. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of experimental psychology. Learning, memory, and cognition*, *25*(4), 997-1010.

American Psychiatric Society. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC.

Anliker, J. (1976). Eye movements: On-line measurement, analysis, and control. In *Eye Movements and Psychological Processes* (pp. 185-199). Hillsdale, NJ: Lawrence Erlbaum Associates.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48-53. doi: 10.1111/j.1467-7687.2007.00563.x.

Avnir, D., Biham, O., Lidar, D., & Malcai, O. (1998). APPLIED MATHEMATICS: Is the Geometry of Nature Fractal? *Science*, *279*(5347), 39-40. doi: 10.1126/science.279.5347.39.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research*, *46*(18), 2824-2833.

Balkenius, C., Eriksson, A. P., & Astrom, K. (2004). Learning in Visual Attention. *Proceedings of LAVS*, *4*.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66-80. doi: 10.1162/jocn.1995.7.1.66.

Baron-Cohen, S. (1995a). *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press.

Baron-Cohen, S. (1995b). *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press.

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., & Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes. *British Journal of Developmental Psychology, 13*(4), 379-398.

Bauke, H. (2007). Parameter estimation for power-law distributions by maximum likelihood methods. *The European Physical Journal B - Condensed Matter and Complex Systems, 58*(2), 167-173. doi: 10.1140/epjb/e2007-00219-y.

Bazell, D., & Desert, F. X. (1988). Fractal structure of interstellar cirrus. *Astrophysical Journal, 333*, 353-358.

Beauchemin, S. S., & Barron, J. L. (1995). The computation of optical flow. *ACM Computing Surveys (CSUR), 27*(3), 433-466.

Belmonte, M. K., Allen, G., Beckel-Mitchener, A., Boulanger, L. M., Carper, R. A., & Webb, S. J. (2004). Autism and Abnormal Development of Brain Connectivity. *Journal of Neuroscience, 24*(42), 9228.

Bhaskar, T., Foo Tun Keat, Ranganath, S., & Venkatesh, Y. (2003). Blink detection and eye tracking for eye localization. In *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region* (Vol. 2, pp. 821-824 Vol.2). doi: 10.1109/TENCON.2003.1273293.

Blake, R., Turner, L. M., Smoski, M. J., Pozdol, S. L., & Stone, W. L. (2003). Visual Recognition of Biological Motion Is Impaired in Children with Autism. *Psychological Science, 14*(2), 151-157. doi: 10.1111/1467-9280.01434.

Boccignone, G., & Ferraro, M. (2004). Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications, 331*(1-2), 207-218.

Boucher, J., & Lewis, V. (1992). Unfamiliar face recognition in relatively able autistic children. *Journal of Child Psychology and Psychiatry*, *33*(5), 843-859.

Braddick, O., Atkinson, J., & Wattam-Bell, J. (2003). Normal and anomalous development of visual motion processing: motion coherence and 'dorsal-stream vulnerability'. *Neuropsychologia*, *41*(13), 1769-1784.

Breazeal, C., & Scassellati, B. (1999). A context-dependent attention system for a social robot. *1999 International Joint Conference on Artificial Intelligence*, 1254–1259.

Brockmann, D., & Geisel, T. (1999). Are human scanpaths Levy flights? *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, *1*.

Brockmann, D., & Geisel, T. (2000). The ecology of gaze shifts. *Neurocomputing*, *32*(33), 643-650.

Bruce, N., & Tsotsos, J. (2005). Saliency Based on Information Maximization. In *Neural Information Processing Systems*. Vancouver, BC.

Brunelli, R., & Poggio, T. (1993). Face recognition: features versus templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *15*(10), 1042-1052.

Bryson, S. E., Rogers, S. J., & Fombonne, E. (2003). Autism Spectrum Disorders: Early Detection, Intervention, Education, and Psychopharmacological Management. *CANADIAN JOURNAL OF PSYCHIATRY*, *48*(8), 506-516.

Burr, D. C., Morrone, M. C., & Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, *371*(6497), 511-513. doi: 10.1038/371511a0.

Burt, P., & Adelson, E. (1983). The Laplacian Pyramid as a Compact Image Code. *Communications, IEEE Transactions on [legacy, pre-1988]*, *31*(4), 532-540.

Burton, G. J., & Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, *26*(1), 157-170.

Buswell, G. T. (1935). *How People Look at Pictures: A Study of the Psychology of Perception in Art.* The University of Chicago press.

Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology, 89*(3), 319-325. doi: 10.1007/s00421-003-0807-5.

Carmi, R., & Itti, L. (2006). Causal saliency effects during natural vision. *Proceedings of the 2006 symposium on Eye tracking research & applications*, 11-18.

Center for Disease Control and Prevention. (2008). Autism Information Center, DD, NCBDDD, CDC. . Retrieved April 29, 2008, from http://www.cdc.gov/ncbddd/autism/index.htm.

Chawarska, K., Klin, A., Paul, R., & Volkmar, F. R. (2007). Autism spectrum disorder in the second year: stability and change in syndrome expression. *Journal of Child Psychology and Psychiatry, 48*(2), 128-138.

Chawarska, K., & Shic, F. Looking but not seeing: Abnormal visual scanning and recognition of faces in 2 and 4-year old children with Autism Spectrum Disorder. Submitted.

Chawarska, K., & Volkmar, F. R. (2007). Impairments in Monkey and Human Face Recognition in 2-Year Toddlers with Autism Spectrum Disorder and Developmental Delay. *Developmental Science, 10*(2), 266-279. doi: 10.1111/j.1467-7687.2006.00543.x.

Chow, G., & Li, X. (1993). Towards a system for automatic facial feature detection. *Pattern Recognition, 26*(12), 1739-1755.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2007). Power-law distributions in empirical data. *0706.1062*. Retrieved July 4, 2008, from http://arxiv.org/abs/0706.1062.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience New York.

Cristinacce, D., & Cootes, T. (2003). Facial feature detection using adaboost with shape constraints. *British Machine Vision Conference, 1*, 231–240.

Crosby, M. E., Iding, M. K., & Chin, D. N. (2001). Visual search and background complexity: Does the forest hide the trees. *User Modeling: Proceedings of the Eighth International Conference, UM2001*, 225–227.

Crowley, J. L., & Berard, F. (1997). Multi-modal tracking of faces for video communications. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 97*, 640–645.

Dakin, S., & Frith, U. (2005). Vagaries of Visual Perception in Autism. *Neuron, 48*(3), 497-507. doi: 10.1016/j.neuron.2005.10.018.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Reviews in Neuroscience, 18*(1), 193-222.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827-1837.

Draper, B., & Lionelle, A. (2003). Evaluation of selective attention under similarity transforms. *Proc. of the Int'l Workshop on Attention and Performance in Computer Vision (WAPCV'03)*, 31–38.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers, 34*, 455-470.

Duchowski, A. T. (2003). *Eye Tracking Methodology: Theory and Practice* (1st ed., p. 252). Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification.* Wiley-Interscience.

Edwards, A. M., Phillips, R. A., Watkins, N. W., Freeman, M. P., Murphy, E. J., Afanasyev, V., et al. (2007). Revisiting Levy flight search patterns of wandering

albatrosses, bumblebees and deer. *Nature*, *449*(7165), 1044-1048. doi: 10.1038/nature06199.

Estes, D. (1994). Young children's understanding of the mind: Imagery, introspection, and some implications. *Journal of Applied Developmental Psychology*, *15*(4), 529-548. doi: 10.1016/0193-3973(94)90021-3.

Falconer, K. (2003). *Fractal Geometry - Mathematical Foundations and Applications* (2nd ed.). Western Sussex, England: John Wiley & Sons, Ltd.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the Internet topology. *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, 251-262.

Fantz, R. L. (1964). Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science*, *146*(3644), 668-670. doi: 10.1126/science.146.3644.668.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, *4*(12), 2379-2394.

Fisher, R. A. (1936). The use of multiple measures in taxonomic problems. *Ann. Eugenics*, *7*, 179-188.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye Movements of Aircraft Pilots During Instrument-Landing Approaches. *Aeronautical Engineering Review*, *9*(2), 24-29.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, *42*(3-4), 143-166.

Frith, C., & Lau, H. (2006). The problem of introspection. *Consciousness and Cognition*, *15*(4), 761-764. doi: 10.1016/j.concog.2006.09.011.

Frith, U. (2003a). *Autism: Explaining the Enigma*. Blackwell Publishers.

Frith, U. (2003b). *Autism: Explaining the Enigma*. Blackwell Publishers.

Frith, U., & Happé, F. (1999). Theory of Mind and Self-Consciousness: What Is It Like to Be Autistic? *Mind & Language*, *14*(1), 82-89. doi: 10.1111/1468-0017.00100.

Geisler, W. S. (2007, December 21). Visual Perception and the Statistical Properties of Natural Scenes. . review-article, . Retrieved July 19, 2008, from http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.psych.58.110405.085632.

Geschwind, D. H., & Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*, *17*(1), 103-111.

Goldberg, J. H., & Schryver, J. C. (1993). Eye-gaze determination of user intent at the computer interface. *Conference: 7. European eye movement conference, Durham (United Kingdom), 31 Aug-3 Sep 1993*.

Goldstein, H. (2002). Communication Intervention for Children with Autism: A Review of Treatment Efficacy. *Journal of Autism and Developmental Disorders*, *32*(5), 373-396. doi: 10.1023/A:1020589821992.

Gordon, H. A. (1981). Errors in Computer Packages. Least Squares Regression Through the Origin. *The Statistician*, *30*(1), 23–29.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*(4), 544-549.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representationof visual salience inmonkey parietal cortex. *Nature*, *391*, 481-484.

Grandin, T. (1992). An inside view of autism. *High functioning individuals with autism*, 105-126.

Green, G., Brennan, L. C., & Fein, D. (2002). Intensive Behavioral Treatment for a Toddler at High Risk for Autism. *Behav Modif*, *26*(1), 69-102. doi: 10.1177/0145445502026001005.

Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., & Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. In

Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on (pp. 222-228).

Guestrin, E. D., & Eizenman, M. (2008). Remote point-of-gaze estimation requiring a single-point calibration for applications with infants. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 267-274). Savannah, Georgia: ACM. doi: 10.1145/1344471.1344531.

Haan, M., Pascalis, O., & Johnson, M. H. (2002). Specialization of Neural Mechanisms Underlying Face Recognition in Human Infants. *Journal of Cognitive Neuroscience, 14*(2), 199-209.

Hain, T. (2008, March 1). Eye movement recording devices. . Retrieved July 10, 2008, from http://www.dizziness-and-balance.com/practice/eyemove.html.

Halit, H., de Haan, M., & Johnson, M. H. (2003). Cortical specialisation for face processing: face-sensitive event-related potential components in 3-and 12-month-old infants. *Neuroimage, 19*(3), 1180-1193.

Happé, F. (1999a). Autism: cognitive deficit or cognitive style? *Trends in Cognitive Sciences, 3*(6), 216-222.

Hayhoe, M. (2000). Vision Using Routines: A Functional Account of Vision. *Visual Cognition, 7*(1-3), 43-64.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188-194.

Heeger, D. J. (1988). Optical flow using spatiotemporal filters. *International Journal of Computer Vision, 1*(4), 279-302.

Heide, W., & Zeec, D. S. (1999). Electrooculography: technical standards and applications. *Recommendations for the Practice of Clinical Neurophysiology: Guidelines of the International Federation of Clinical Neurophysiology.*

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, 537–562.

Henderson, J. M., & Hollingworth, A. (1999). High-Level Scene Perception. *Annual Reviews in Psychology*, *50*(1), 243-271.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, *7*(11), 498-504. doi: 10.1016/j.tics.2003.09.006.

Hjelmas, E., & Low, B. K. (2001). Face detection: A survey. *Computer Vision and Image Understanding*, *83*(3), 236-274.

Hobson, R. P., Ouston, J., & Lee, A. (1988). What's in a face? The case of autism. *Br J Psychol*, *79*(Pt 4), 441-53.

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*, 80-84.

Hoffman, J. E., & Subramaniam, B. (1995). The Role of Visual Attention in Saccadic Eye Movements. *Perception and Psychophysics*, *57*(6), 787-795.

Hornof A.J., & Halverson T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers*, *34*, 592-604.

Huey, E. B. (1898). Preliminary Experiments in the Physiology and Psychology of Reading. *The American Journal of Psychology*, *9*(4), 575-586.

Inhoff, A. W., Pollatsek, A., Posner, M. I., & Rayner, K. (1989). Covert attention and eye movements during reading. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *41*(1), 63-89. doi: 2710940.

Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. *Eye guidance in reading and scene perception*, 1–28.

Irwin, D. (2004). Fixation Location and Fixation Duration as Indices of Cognitive Processing. In *The Interface of Language, Vision, and Action: Eye Movements and the Visual World.* Psychology Press.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093-1123.

Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition, 14*(4), 959-984.

Itti, L. (2008). iLab Neuromorphic Vision C++ Toolkit (iNVT). . Retrieved July 20, 2008, from http://ilab.usc.edu/toolkit/.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *Advances in Neural Information Processing Systems, 19*, 1-8.

Itti, L., Dhavale, N., & Pighin, F. (2003). Realistic avatar eye and head animation using a neurobiological model of visual attention. *Proceedings of SPIE, 5200*, 64-78.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis . *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20*(11), 1254-1259. doi: 10.1109/34.730558.

Itti, L., & Koch, C. (1999). Comparison of feature combination strategies for saliency-based visual attention systems. In *Human Vision and Electronic Imaging IV* (Vol. 3644, pp. 473-482). San Jose, CA, USA: SPIE. Retrieved from http://link.aip.org/link/?PSI/3644/473/1.

Itti, L., Rees, G., & Tsotsos, J. K. (2005). *Neurobiology of Attention.* Academic Press.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises (Section commentary). In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573-605). Oxford, England: Elsevier Science BV.

Jain, R., Kasturi, R., & Schunck, B. G. (1995). *Machine vision.*

James, A., & Plank, M. J. (2007). On fitting power laws to ecological data. *Arxiv preprint arXiv:0712.0613*.

Jeng, S. H., Liao, H. Y. M., Han, C. C., Chern, M. Y., & Liu, Y. T. (1998). Facial feature detection using geometrical face model: An efficient approach. *Pattern Recognition, 31*(3), 273-282.

Just, M. A., & Carpenter, P. A. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review, 87*(4), 329-54.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311.

Karmiloff-Smith, A. (2007). Atypical epigenesis. *Developmental Science, 10*(1), 84-88.

Karsh, R., & Breitenbach, F. W. (1983). Looking at looking: The amorphous fixation measure. *Eye Movements and Psychological Functions: International Views*, 53-64.

Kawato, S., & Tetsutani, N. (2004). Detection and tracking of eyes for gaze-camera control. *Image and Vision Computing, 22*(12), 1031-1038.

Khan, I. R., & Ohba, R. (1999). Closed-form expressions for the finite difference approximations of first and higher derivatives based on Taylor series. *Journal of Computational and Applied Mathematics, 107*(2), 179-193. doi: 10.1016/S0377-0427(99)00088-6.

Klin, A., Jones, W., Schultz, R., & Volkmar, F. R. (2003). The Enactive Mind, or from Actions to Cognition: Lessons from Autism. *Philosophical Transactions: Biological Sciences, 358*(1430), 345-360.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002a). Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism. *Archives of General Psychiatry, 59*(9), 809.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002b). Defining and Quantifying the Social Phenotype in Autism. *Am J Psychiatry, 159*(6), 895-908. doi: 10.1176/appi.ajp.159.6.895.

Klin, A., Lin, D., Gorrindo, P., & Jones, W. Two-year-olds with autism fail to orient towards human biological motion but attend instead to non-social, physical contingencies. .

Klin, A., Sparrow, S. S., de Bildt, A., Cicchetti, D. V., Cohen, D. J., & Volkmar, F. R. (1999). A Normed Study of Face Recognition in Autism and Related Disorders. *Journal of Autism and Developmental Disorders, 29*(6), 499-508.

Klinkenberg, B. (1994). A review of methods used to determine the fractal dimension of linear features. *Mathematical Geology, 26*(1), 23-46. doi: 10.1007/BF02065874.

Koch, C. (1984). A theoretical analysis of the electrical properties of an X-cell in the cat's LGN: does the spine-triad circuit subserve selective visual attention. *Artif. Intell. Memo, 787.*

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol, 4*(4), 219-27.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897-1916.

Kruizinga, A., Mulder, B., & de Waard, D. (2006). Eye scan patterns in a simulated ambulance dispatcher's task. In D. de Waard, K. Brookhuis, & A. Toffetti (Eds.), *Developments in Human Factors in Transportation, Design, and Evaluation* (pp. 305-317). Orbassano, Italy: Shaker Publishing.

Kustov, A. A., & Lee Robinson, D. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*(6604), 74-77.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research, 41*(25-26), 3559-3565.

Lawson, W. (2001). *Understanding and Working with the Spectrum of Autism: An Insider's View*. Jessica Kingsley Publishers.

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, *2*, 375-381.

Leigh, R. J., & Zee, D. S. (2006). *The Neurology of Eye Movements: Book-and-DVD Package* (4th ed., p. 776). Oxford University Press, USA.

Lewis, J. (1995). Fast normalized cross-correlation. *Vision Interface*, 120—123.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9-16.

Liang, J., Moshel, S., Zivotofsky, A. Z., Caspi, A., Engbert, R., Kliegl, R., et al. (2005). Scaling of horizontal and vertical fixational eye movements. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, *71*(3), 031909-6.

Lin, D., Jones, W., Shic, F., Knoch, K., Shultz, S., & Klin, A. (2007). Effects of Audio-Visual Synchrony on the Viewing Patterns of Children with Autism. In *Proceedings of the 6th Annual International Meeting for Autism Research (IMFAR 2007)*. Seattle, Washington.

Lord, C. (2002). *Autism Diagnostic Observation Schedule: ADOS: Manual*. Western Psychological Services.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, *55*(1), 3-9.

Lowe, D. G. (1999). Object Recognition from Local Scale-Invariant Features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu* (pp. 1150-1157). Retrieved July 30, 2008, from http://citeseer.ist.psu.edu/lowe99object.html.

Lundqvist, D., Flykt, A., & Ohman, A. (1998). The Karolinska Directed Emotional Faces. *Pictoral face set available from the Department of Neurosciences, Karolinska Hospital, Stockholm, Sweden.*

Mandelbrot, B. B. (1967). How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension. *Science*, *156*(3775), 636.

Mandelbrot, B. B. (1975). Stochastic Models for the Earth's Relief, the Shape and the Fractal Dimension of the Coastlines, and the Number-Area Rule for Islands. *Proceedings of the National Academy of Sciences of the United States of America*, *72*(10), 3825-3828.

Mandelbrot, B. B. (1999). A fractal walk down Wall Street. *Scientific American*, 70-73.

Mandelbrot, B. B., Passoja, D. E., & Paullay, A. J. (1984). Fractal character of fracture surfaces of metals. *Nature*, *308*(5961), 721-722.

Mayes, L. C., & Cohen, D. J. (1994). Experiencing Self and Others: Contributions from Studies of Autism to the Psychoanalytic Theory of Social Development. *Journal of the American Psychoanalytic Association*, *42*, 191-218.

Mayes, L. C., Bornstein, M. H., Chawarska, K., & Granger, R. H. (1995). Information Processing and Developmental Assessments in 3-Month-Old Infants Exposed Prenatally to Cocaine. *Pediatrics*, *95*(4), 539-545.

Mazer, J. A., & Gallant, J. L. (2003). Goal-Related Activity in V4 during Free Viewing Visual Search Evidence for a Ventral Stream Visual Salience Map. *Neuron*, *40*(6), 1241-1250.

McEachin, J. J., Smith, T., & Lovaas, O. I. (2001). Long-term outcome for children with autism who received early intensive behavioral treatment. *The Science of Mental Health.*

Meltzoff, A., & Moore, M. (1977). Imitation of Facial and Manual Gestures by Human Neonates. *Science*, *198*(4312), 75-78.

Milne, E., Swettenham, J., Hansen, P., Campbell, R., Jeffries, H., & Plaisted, K. (2002). High Motion Coherence Thresholds in Children with Autism. *Journal of Child Psychology and Psychiatry*, *43*(2), 255-263. doi: 10.1111/1469-7610.00018.

Morimoto, C., Amir, A., & Flickner, M. (2002). Detecting eye position and gaze from a single camera and 2 light sources. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on* (Vol. 4, pp. 314-317 vol.4). doi: 10.1109/ICPR.2002.1047459.

Mullen, E. M. (1995). Mullen Scales of Early Learning. *Circle Pines, MN: American Guidance Service.*

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387-391. doi: 10.1038/nature03390.

Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development*, *10*(1-2), 3-18. doi: 10.1002/icd.239.

Neumann, D., Spezio, M. L., Piven, J., & Adolphs, R. (2006). Looking you in the mouth: abnormal gaze in autism resulting from impaired top-down modulation of visual attention. *Social Cognitive and Affective Neuroscience*, *1*(3), 194.

Niebur, E., Itti, L., & Koch, C. (1995). Modeling the "where" visual pathway. In *Proceedings of the 2nd Joint Symposium on Neural Computation* (Vol. 5, pp. 26-35). Institute for Neural Computationa, La Jolla, San Diego, CA.

Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modeling the where pathway. *Advances in Neural Information Processing Systems*, *8*, 802-808.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

Noton, D., & Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception. *Science*, New Series., *171*(3968), 308-311.

Ouerhani, N., von Wartburg, R., Hugli, H., & Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, *3*(1), 13-24.

Overgaard, M. (2006). Introspection in Science. *Consciousness and Cognition*, *15*(4), 629-633. doi: 10.1016/j.concog.2006.10.004.

Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive Function Deficits in High-Functioning Autistic Individuals: Relationship to Theory of Mind. *Journal of Child Psychology and Psychiatry*, *32*(7), 1081-1105.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107-123.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125-154.

Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., & Piven, J. (2002). Visual Scanning of Faces in Autism. *Journal of Autism and Developmental Disorders*, *32*(4), 249-261.

Petersen, S. E., Robinson, D. L., & Morris, J. D. (1987). Contributions of the pulvinar to visual spatial attention. *Neuropsychologia*, *25*(1A), 97-105.

Pierce, K., Muller, R. A., Ambrose, J., Allen, G., & Courchesne, E. (2001). Face processing occurs outside the fusiform'face area'in autism: evidence from functional MRI. *Brain*, *124*(10), 2059.

Plank, M. J., & James, A. (2008). Optimal foraging: Lévy pattern or process? *Journal of the Royal Society, Interface / the Royal Society*. doi: J3RU04H1U2773U7P.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3-25.

Posner, M. I., & Petersen, S. E. (1990). The Attention System of the Human Brain. *Annual Reviews in Neuroscience*, *13*(1), 25-42.

Privitera, C. M., & Stark, L. W. (2000). Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 970-982.

Radach, R., & Kennedy, A. (2004). Theoretical perspectives on eye movements in reading: Past controversies, current issues, and an agenda for future research. *Eye Movements and Information Processing During Reading*, *16*(1/2), 3-26.

Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. C. (2005). Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A*, *22*(10), 2039-2049.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372-422.

Reinagel, P., & Zador, A. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*(4), 341-350.

Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, *7*(3), 6.

Richmond, J., Colombo, M., & Hayne, H. (2007). Interpreting visual preferences in the visual paired-comparison task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 823-831.

Richmond, J., Sowerby, P., Colombo, M., & Hayne, H. (2004). The effect of familiarization time, retention interval, and context change on adults' performance in the visual paired-comparison task. *Developmental Psychobiology*, *44*(2), 146-155. doi: 10.1002/dev.10161.

Rinehart, N. J., Bradshaw, J. L., Moss, S. A., Brereton, A. V., & Tonge, B. J. (2000). Atypical Interference of Local Detail on Global Processing in High- Functioning Autism and Asperger's Disorder. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(06), 769-778. doi: null.

Roberts, D., Shelhamer, M., & Wong, A. (2008). A new "wireless" search-coil system. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 197-204). Savannah, Georgia: ACM. doi: 10.1145/1344471.1344519.

Robinson, D. L., & Petersen, S. E. (1992). The pulvinar and visual salience. *Trends Neurosci*, *15*(4), 127-32.

Rodríguez-Iturbe, I., & Rinaldo, A. (1997). *Fractal River Basins: Chance and Self-Organization*. Cambridge University Press.

Rogers, S. J. (1998). Empirically supported comprehensive treatments for young children with autism. *Journal of Clinical Child Psychology*, *27*(2), 168-179.

Rowley, H. A., Baluja, S., & Kanade, T. (1998). Neural network-based face detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *20*(1), 23-38.

Saber, E., & Murat Tekalp, A. (1998). Frontal-view face detection and facial feature extraction using color, shape and symmetry based cost functions. *Pattern Recognition Letters*, *19*(8), 669-680.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the 2000 symposium on Eye tracking research & applications*, 71-78.

Santella, A., & Decarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. *Eye Tracking Research & Application: Proceedings of the 2004 symposium on Eye tracking research & applications*, *22*(24), 27-34.

Schneiderman, H., & Kanade, T. (2000). Statistical method for 3 D object detection applied to faces and cars. *PROC IEEE COMPUT SOC CONF COMPUT VISION PATTERN RECOGNIT*, *1*, 746-751.

Schultz, R. T., Gauthier, I., Klin, A., Fulbright, R. K., Anderson, A. W., Volkmar, F. R., et al. (2000). Abnormal Ventral Temporal Cortical Activity During Face Discrimination Among Individuals With Autism and Asperger Syndrome. *Archives of General Psychiatry*, *57*(4), 331.

Schuster, F. L., & Levandowsky, M. (1996). Chemosensory responses of Acanthamoeba castellanii: visual analysis of random movement and responses to chemical signals. *The Journal of Eukaryotic Microbiology*, *43*(2), 150-8. doi: 8720945.

SensoMotoric Instruments (SMI). (2006, November). iView X: Eye and Gaze Tracker. . Retrieved July 11, 2008, from http://web.archive.org/web/20061013204944/www.smi.de/iv/index.html.

Shalizi, C. So You Think You Have a Power Law — Well Isn't That Special? . Retrieved July 4, 2008, from http://cscs.umich.edu/~crshalizi/weblog/491.html.

Sheinkopf, S. J., & Siegel, B. (1998). Home-Based Behavioral Treatment of Young Children with Autism. *Journal of Autism and Developmental Disorders*, *28*(1), 15-23. doi: 10.1023/A:1026054701472.

Shelhamer, M. (2005a). Sequences of Predictive Saccades Are Correlated Over a Span of ~2 s and Produce a Fractal Time Series. *J Neurophysiol*, *93*(4), 2002-2011. doi: 10.1152/jn.00800.2004.

Shelhamer, M. (2005b). Phase transition between reactive and predictive eye movements is confirmed with nonlinear forecasting and surrogates. *Neurocomputing*, *65-66*, 769-776. doi: 10.1016/j.neucom.2004.10.073.

Shelhamer, M. (2005c). Sequences of predictive eye movements form a fractional Brownian series – implications for self-organized criticality in the oculomotor system. *Biological Cybernetics*, *93*(1), 43-53. doi: 10.1007/s00422-005-0584-9.

Shelhamer, M., & Joiner, W. M. (2003). Saccades Exhibit Abrupt Transition Between Reactive and Predictive, Predictive Saccade Sequences Have Long-Term Correlations. *J Neurophysiol*, *90*(4), 2763-2769. doi: 10.1152/jn.00478.2003.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology Section A*, *38*(3), 475-491.

Shic, F., Chawarska, K., Bradshaw, J., & Scassellati, B. (2008). Autism, Eye-Tracking, Entropy. In *7th Annual IEEE Conference on Development and Learning.* Monterey, CA.

Shic, F., Chawarska, K., Lin, D., & Scassellati, B. (2007). Measuring context: The gaze patterns of children with autism evaluated from the bottom-up. In *Development and Learning, 2007. ICDL IEEE 6th International Conference on* (pp. 70-75).

Shic, F., Chawarska, K., Lin, D., & Scassellati, B. (2008). The Computational Modeling of Perceptual Biases of Children with ASD in Naturalistic Settings. In

Proceedings of the 7th Annual International Meeting for Autism Research (IMFAR 2008). London, UK.

Shic, F., Chawarska, K., & Scassellati, B. (2008a). The incomplete fixation measure. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (pp. 111-114). Savannah, Georgia: ACM. doi: 10.1145/1344471.1344500.

Shic, F., Chawarska, K., & Scassellati, B. (2008b). The Amorphous Fixation Measure Revisited: with Applications to Autism. In *30th Annual Meeting of the Cognitive Science Society*. Washington, DC.

Shic, F., Chawarska, K., Zucker, S. W., & Scassellati, B. (2008). Fractals from Fixations. In *Proceedings of the 41st Annual Society for Mathematical Psychology Conference*. Washington, DC.

Shic, F., Jones, W., Klin, A., & Scassellati, B. (2006). Swimming in the Underlying Stream: Computational Models of Gaze in a Comparative Behavioral Analysis of Autism. In *28th Annual Conference of the Cognitive Science Society*.

Shic, F., & Scassellati, B. (2007). A Behavioral Analysis of Computational Models of Visual Attention. *International Journal of Computer Vision*, *73*(2), 159-177. doi: 10.1007/s11263-006-9784-6.

Sims, D. W., Righton, D., & Pitchford, J. W. (2007). Minimizing Errors in Identifying Levy Flight Behaviour of Organisms. *Journal of Animal Ecology*, *76*(2), 222-229. doi: 10.1111/j.1365-2656.2006.01208.x.

Singer-Vine, J. (2008, July). New Ways to Diagnose Autism Earlier - WSJ.com. *The Wall Street Journal*. Retrieved July 11, 2008, from http://online.wsj.com/article/SB121545978096433273.html.

Smith, T. (1999). Outcome of Early Intervention for Children With Autism. *Clinical Psychology: Science and Practice*, *6*(1), 33-49. doi: 10.1093/clipsy.6.1.33.

Tantam, D., Monaghan, L., Nicholson, H., & Stirling, J. (1989). Autistic Children's Ability to Interpret Faces: a Research Note. *Journal of Child Psychology and Psychiatry*, *30*(4), 623-630. doi: 10.1111/j.1469-7610.1989.tb00274.x.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research, 45*(5), 643-659. doi: 10.1016/j.visres.2004.09.017.

Taylor, R. P., Micolich, A. P., & Jonas, D. (1999). Fractal analysis of Pollock's drip paintings. *Nature, 399*(6735), 422. doi: 10.1038/20833.

Tolhurst, D. J., Tadmor, Y., & Chao, T. (1992). Amplitude spectra of natural images. *Ophthalmic and Physiological Optics, 12*(2), 229-232. doi: 10.1111/j.1475-1313.1992.tb00296.x.

Torralba, A. (2003). Modeling global scene factors in attention. *Journal of the Optical Society of America A, 20*(7), 1407-1418.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems, 14*(3), 391-412.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136.

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Current Opinion in Neurobiology, 13*(4), 428-432.

Tsotsos, J. K. (1988). A 'complexity level' analysis of immediate vision. *International Journal of Computer Vision, 1*(4), 303-320. doi: 10.1007/BF00133569.

Tsotsos, J. K., Culhane, S. M., Kei Wai, W. Y., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence, 78*(1-2), 507-545.

Tsotsos, J. K., Liu, Y., Martinez-Trujillo, J. C., Pomplun, M., Simine, E., & Zhou, K. (2005). Attending to visual motion. *Computer Vision and Image Understanding, 100*(1-2), 3-40.

Turano, K. A., Geruschat, D. R., & Baker, F. H. (2003). Oculomotor strategies for the direction of gaze tested with a real-world activity. *Vision Research, 43*(3), 333-346.

UK postal areas. . (2008). Retrieved July 18, 2008, from http://www.werelate.org/wiki/Image:UK_postal_areas.png.

Urruty, T., Lew, S., Djeraba, C., & Simovici, D. (2007). Detecting Eye Fixations by Projection Clustering. In *Image Analysis and Processing Workshops, 2007. ICIAPW 2007. 14th International Conference on* (pp. 45-50). doi: 10.1109/ICIAPW.2007.22.

Valenza, E., Simion, F., Cassia, V. M., & Umilta, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 892-903.

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, *57*(2), 137-154.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., & Stanley, H. E. (1996). Levy flight search patterns of wandering albatrosses. *Nature*, *381*(6581), 413-415. doi: 10.1038/381413a0.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G. E., Raposo, E. P., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*(6756), 911-914. doi: 10.1038/44831.

White, E. P., Enquist, B. J., & Green, J. L. (2008). ON ESTIMATING THE EXPONENT OF POWER-LAW FREQUENCY DISTRIBUTIONS. *Ecology*, *89*(4), 905-912 . doi: 10.1890/07-1288.1.

Widdel, H. (1984). Operational problems in analysing eye movements. *Theoretical and Applied Aspects of Eye Movement Research*, 21-29.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202-238.

Wolfe, J. M., & Gancarz, G. (1996). Guided Search 3.0: A model of visual search catches up with Jay Enoch 40 years later. *Basic and clinical applications of vision science*, 189–192.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nat Rev Neurosci, 5*(6), 495-501.

Wong, P., Howard, J., & Lin, J. S. (1986). Surface Roughening and the Fractal Nature of Rocks. *Physical Review Letters, 57*(5), 637-640.

Yarbus, A. L. (1967). *Eye Movements and Vision.* Plenum Press.

Yee, C., & Walther, D. (2002). *Motion detection for bottom-up visual attention.* tech. rep., SURF/CNS, California Institute of Technology, 2002.

Yoo, D. H., Kim, J. H., Lee, B. R., & Chung, M. J. (2002). Non-contact eye gaze tracking system by mapping of corneal reflections. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* (pp. 94-99). doi: 10.1109/AFGR.2002.1004139.

Zaharescu, A., Rothenstein, A. L., & Tsotsos, J. K. (2005). Towards a Biologically Plausible Active Visual Search Model. In *Attention and Performance in Computational Vision* (pp. 133-147). Retrieved July 20, 2008, from http://www.springerlink.com/content/c4pj38d2hrp39xey.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry, 39*(4), 561-77. doi: 8472349.