Attention-Aware Rendering, Mobile Graphics and Games

Ann McNamara, Texas A&M University
Katerina Mania, Technical University of Crete
George Koulieris, Technical University of Crete
Laurent Itti, University of Southern California

Siggraph 2014 Course Notes
Vancouver BC 2014

ACM 978-1-4503-2962-0/14/08

Abstract

Rendering and design efficiency become critical when deploying computer graphics to mobile devices and games. This course addresses novel approaches to leverage models of visual attention, based on low and high level scene features, to propel attention-aware rendering computation. The result: perceptually-optimized scalable algorithms for mobile platforms and game design.

Recent research on low-level attention algorithms whose architecture and function is closely inspired from biological brains will be reviewed (Baldi & Itti, 2010). High level saliency instigated by cognitive information such as scene context and topology will also be discussed (Zotos, Mania, & Mourkoussis, 2009; ?, ?). Prediction of attention can significantly improve many aspects of computer graphics and games. Image synthesis can be accelerated by reducing computation on non-attended scene regions. Subtle gaze manipulation (Bailey, McNamara, Sudarsanam, & Grimm, 2009a) improves mammography training and spatial recall as well as dynamically influence interactive neuroscientific protocols in fMRI.

The course will showcase through interactive content how employing a visual attention model leads to efficient game design. Many games rely on search or target detection tasks to solve riddles or find game objects. Adjusting the difficulty of a level could be facilitated by relocating objects estimated to attract attention. By incorporating visual attention estimation, perceptually optimized renderers for mobile platforms can dynamically ignore perceptually non-important details. Modern video games and interactive applications are simultaneously deployed for computers, consoles and mobile devices of diverse computational power. Complex effects including but not limited to complex refraction and subsurface scattering that are standard in desktop computers, could also scale efficiently in portable devices if attention-aware. Examples such as the integration of a high level saliency model in a LOD manager, enabling complex effects in low-power devices by applying them in regions expected to be attended to, are going to be demonstrated.

This course delivers a cutting-edge overview of attention models and their application in rendering, mobile technology and gaming platforms.

About the Lecturers

Ann McNamara

Dr. Ann McNamara is an assistant professor in the Department of Visualization at Texas A&M University. She is the recipient of an NSF CAREER award. Her research focuses on the advancement of computer graphics and scientific visualization through novel approaches for optimizing an individual's experience when creating, viewing and interacting with real, augmented and virtual spaces. She investigates new ways to exploit knowledge of human visual perception to produce high quality computer graphics and animations more efficiently. Ann was Siggraph 2011 & Siggraph 2012 Courses Chair. She is Siggraph General Submissions Chair for Siggraph 2014 & Siggraph 2015.

Katerina Mania

Katerina Mania completed a B.Sc. in Mathematics, University of Crete, an M.Sc./Ph.D in Computer Science, University of Bristol funded by HP Labs. She worked at HP Labs as a researcher and as faculty in the University of Sussex. Katerina spent her sabbaticals at NASA Ames Research Centre (Advanced Displays/Spatial Perception Laboratory) in 2003 and at Brighton and Sussex Medical School in 2013. She is now an Associate Professor at the Technical University of Crete, an Associate Editor of ACM Transactions on Applied Perception and Presence Teleoperators and Virtual Environments and in the steering committee of the Symposium on Applied Perception.

George Alex Koulieris

George Alex Koulieris completed a B.Sc. in Computer Science, University of Athens, and an MSc in Computer Science, University of Economics and Business, Athens. He is a PhD candidate experienced in interdisciplinary research at the Dept. of Electronic & Computer Engineering of the Technical University of Crete working on attention modeling optimized by cognitive psychology principles. For his BSc thesis titled Efficient Soft Shadow Mapping, he worked at INRIA Sophia Antipolis in France (REVES group), under the supervision of Dr. George Drettakis and Prof. Theoharis Theoharis. His Masters thesis was titled Medical MRI Volume Data Processing and Visualization.

Laurent Itti

Laurent Itti received his M.S. in Image Processing from Ecole Nationale Superieure des Telecommunications (Paris, France) in 1994, and his Ph.D. in Computation and Neural Systems from Caltech (Pasadena, California) in 2000. He has been an Assistant, Associate, and now Full Professor of Computer Science, Psychology, and Neuroscience, University of Southern California. Dr. Itti's research interests are in biologically-inspired computational vision, visual attention, scene understanding, eye movements, and surprise. This basic research has technological applications to, among others, video compression, target detection, robotics. Dr. Itti has co-authored over 130 publications, three patents, and several open-source neuropmorphic vision software toolkits.

Course Overview

15 minutes: Welcome, Introductions, Background and Motivation

Ann McNamara Welcome, overview of course & motivation for computer graphics. Speaker Introductions.

45 minutes: Visual-Attention Modeling

Laurent Itti Visual attention on Complex Interactive Environments

30 minutes: Directing Attention

Ann McNamara Directing attention in art, augmented reality, medical applications and tiled displays

45 minutes: High level saliency in games, rendering and mobile platforms

Katerina Mania & George Koulieris Prediction of saliency and optimization in games, rendering and mobile platforms

30 minutes: Neuromorphic models of vision and applications in graphics

Laurent Itti Neurobiology of attention

15 minutes: Eye movements and attention of neuro-scientific protocols in fMRI

Katerina Mania & George Koulieris In collaboration with Brighton and Sussex Medical School

10 minutes: A summary of related cutting edge perceptual research selected from SAP 2014

Ann McNamara, Katerina Mania & George Alex Koulieris

10 minutes: Conclusion, Questions & Answers

All Wrap up, review, questions and discussion.

Contents

Chapter 1

Introduction

This course presents timely, relevant examples on how researchers have leveraged prediction of attention based on low and high level features for optimization of rendering algorithms and rendering on mobile platforms and to better guide design of games. Each presentation will provide references and short overviews of cutting-edge current research pertaining to that area. We will ensure that the most up-to-date research examples are presented by sourcing information from recent perception, vision and graphics conferences paying particular attention at work presented at the Symposium on Applied Perception in Graphics and Visualization 2014, co-located with ACM SIGGRAPH 2014. In particular, we will review recent research on low-level computational neuroscience algorithms whose architecture and function is closely inspired from biological brains [Baldi and Itti 2010]. High Level Saliency instigated by cognitive information such as scene context and topology will be discussed next [Henderson et al. 1999].

1.1 Summary Statement

This course presents timely examples on how researchers have leveraged prediction of attention based on low and high level scene features for optimization of rendering algorithms, rendering on mobile platforms and to guide design of games. Each presentation will provide references and overviews of cutting-edge research pertaining that area.

1.2 Short Overview

The prediction of attention can significantly improve many aspects of computer graphics and games. For example, image synthesis can be accelerated by reducing computation on non-attended scene regions; attention can also be used to improve LOD. Another interesting case is computer game design. Many game genres rely on a search or target detection task to solve riddles or find game objects. If attention can be automatically predicted, several tasks in game design would be simplified. For example adjusting the difficulty of a game level could be facilitated by relocating objects estimated to attract attention.
Modern video games and interactive applications get simultaneously deployed for computers, consoles and mobile devices. These platforms are extremely diverse in terms of computing power. Modern materials and effects such as complex refraction with chromatic aberration and subsurface scattering that are considered standard in high-end desktop computers, do not scale well in portable devices. Hardware restrictions prohibit the use of *necessary for the game aesthetics* effects that demand multiple texture fetches and many arithmetic/logic operations. By exploiting cognitive effects in human attention we can design a perceptually optimized renderer that takes into account the characteristics of attention deployment and save computational time by removing perceptually non-important details. Integration of a high level saliency model in a level of detail manager enables the usage of complex effects in low-power devices by applying them sparingly only in regions that are expected to be attended.We will ensure that the most up-to-date research examples

are presented by sourcing information from recent perception, vision, neuroscience and graphics conferences paying particular attention at work presented at the Symposium on Applied Perception in Graphics and Visualization 2014, co-located with ACM SIGGRAPH 2014.

1.3 Course Objectives

Efficient rendering for computer graphics continues to be an active area of research and development. This research is driven by the desire to accurately represent virtual environments, while keeping computational costs to a minimum. These costs become even more critical when deploying graphics to mobile devices and gaming paradigms. To gain efficiencies many researchers employ computational models of visual attention to propel computation into those areas of a scene most noticeable by human observers. This course delivers a comprehensive overview of attention models and shows their usefulness in applications including rendering in computer graphics, mobile technology and gaming platforms.

Chapter 2

Computational Models of Attention

2.1 Abstract

This chapter reviews recent progress in computational modeling of visual attention. We start with early concepts and models, which have emphasized stimulus-driven guidance of attention towards salient objects in the visual world. We then present a taxonomy of the many different approaches which have emerged in recent research efforts. We then turn to the more complex problem of modeling top-down, task- and goal-driven influences on attention. While early top-down models have been more qualitative in nature, we describe several recent fully computational approaches that address top-down biasing in space, over features, and towards objects. We finally provide an outlook and describe promising future research directions.

2.2 Preliminaries and definitions

Computational models of visual attention have become popular over the past decade, we believe primarily for two reasons: First, models make testable predictions that can be explored by experimentalists as well as theoreticians; second, models have practical and technological applications of interest to the applied science and engineering communities. In this chapter, we take a critical look at recent attention modeling efforts. We focus on *computational models of attention* as defined by Tsotsos & Rothenstein (Tsotsos & Rothenstein, 2011): Models which can process any visual stimulus (typically, an image or video clip), which can possibly also be given some task definition, and which make predictions that can be compared to human or animal behavioral or physiological responses elicited by the same stimulus and task. Thus, we here place less emphasis on abstract models, phenomenological models, purely data-driven fitting or extrapolation models, or models specifically designed for a single task or for a restricted class of stimuli. For theoretical models, we refer the reader to a number of previous reviews that address attention theories and models more generally (Itti & Koch, 2001b; Paletta, Rome, & Buxton, 2005; Frintrop, Rome, & Christensen, 2010a; Rothenstein & Tsotsos, 2008; Gottlieb & Balan, 2010; Toet, 2011; ?, ?).

To frame our narrative, we embrace a number of notions that have been popularized in the field, even though many of them are known to only represent coarse approximations to biophysical or psychological phenomena. These include the *attention spotlight* metaphor (Crick, 1984), the role of focal attention in *binding* features into coherent representations (A. M. Treisman & Gelade, 1980a), and the notions of an *attention bottleneck*, a *nexus*, and an *attention hand* as embodiments of the attentional selection process (R. Rensink, 2000; Navalpakkam & Itti, 2005). Further, we cast the problem of modeling attention computationally as comprising at least three facets: *guidance* (that is, which computations are involved in deciding where or what to attend to next?), *selection* (how is attended information segregated out of other incoming sensory information?), and *enhancement* (how is the information selected by attention processed differently than non-selected information?). While different theories and models have addressed all three aspects, most computational models as defined above have focused on the initial and primordial problem of guidance. Thus guidance is our primary focus, and we refer the reader to previous reviews on selection and enhancement (Allport, Meyer, & Kornblum, 1993; Desimone & Duncan, 1995; Reynolds & Desimone, 1999; Driver, 2001;

Robertson et al., 2003; Carrasco, 2011). Note that guidance of attention is often thought of as involving *pre-attentive* computations to attract the focus of attention to the next most behaviorally relevant location (hence, attention guidance models might — strictly speaking — be considered pre-attention rather than attention models).

We explore models for exogenous (or bottom-up, stimulus-driven) attention guidance as well as for endogenous (or top-down, context-driven, or goal-driven) attention guidance. Bottom-up models process sensory information primarily in a feed-forward manner, typically applying successive transformations to visual features received over the entire visual field, so as to highlight those locations which contain the most interesting, important, conspicuous, or so-called *salient* information (Koch & Ullman, 1985; Itti & Koch, 2001b). Many, but not all, of these bottom-up models embrace the concept of a topographic saliency map, which is a spatial map where the map value at every location directly represents visual salience, abstracted from the details of why a location is salient or not (Koch & Ullman, 1985). Under the saliency map hypothesis, the task of a computational model is then to transform an image into its spatially corresponding saliency map, possibly also taking into account temporal relations between successive video frames of a movie (Itti, Koch, & Niebur, 1998). Many models thus attempt to provide an operational definition of salience in terms of some image transform or some importance operator that can be applied to an image and that directly returns salience at every location, as we further examine below.

By far, the bottom-up, stimulus-driven models of attention have been more developed, probably because they are task-free, and thus often require no learning, training, or tuning to open-ended task or contextual information. This makes the definition of a purely bottom-up importance operator tractable. Another attractive aspect of bottom-up models — and especially saliency map models — is that, once implemented, they can easily be applied to any image and yield some output that can be tested against human or animal experimental data. Thus far, the most widely used test to validate predictions of attention models has been direct comparison between model output and eye movements recorded from humans or animals watching the same stimuli as given to the models (see, e.g., among many others, (Parkhurst, Law, & Niebur, 2002; Itti, 2006; Le Meur, Le Callet, Barba, & Thoreau, 2006)). Recently, several standard benchmark image and video datasets with corresponding eye movement recordings from pools of observers have been adopted, which greatly facilitates quantitative comparisons of computational models (Carmi & Itti, 2006; Judd, Ehinger, Durand, & Torralba, 2009a; Toet, 2011; J. Li, Tian, Huang, & Gao, 2011; Borji, Sihite, & Itti, 2012b). It is important to note, however, that several alternative measures have been employed as well (e.g., mouse clicks, search efficiency, reaction time, optimal information gain, scanpath similarity; see (M. Eckstein, Peterson, Pham, & Droll, 2009; ?, ?)). One caveat of metrics that compare model predictions to eye movements is that the distinction between covert and overt attention is seldom explicitly addressed by computational models: models usually produce as their final result a saliency map without further concern of how such map may give rise to an eye movement scanpath (Noton & Stark, 1971) (but see (Itti, Dhavale, & Pighin, 2003; X. Ma & Deng, 2009; X. Sun, Yao, & Ji, 2012) and active vision/robotics systems like (Orabona, Metta, & Sandini, 2005; Frintrop, 2006; Belardinelli, Pirri, & Carbone, 2006; Ajallooeian, Borji, Araabi, Ahmadabadi, & Moradi, 2009)). Similarly, biases in datasets, models, and/or behavior may affect the comparison results (Tatler & Vincent, 2009; Tseng, Carmi, Cameron, Munoz, & Itti, 2009; Bonev, Chuang, & Escolano, 2012). While these are important for specialist audiences, here we should just remember that quantitative model evaluation metrics based on eye movements exist and are quite well established, although they remain approximate and should be used with caution (Tatler, Baddeley, & Gilchrist, 2005). Beyond metric issues, one last important consideration to ensure a valid direct comparison between model and eye movements is that conditions should be exactly matched between the model run and the experimental participants, which has not always been the case in previous work, as we discuss below.

While new bottom-up attention models are constantly proposed which rely on novel ways of analyzing images to determine the most interesting or salient locations (we list over 50 of them below), several research efforts have also started to address the more complex problem of building top-down models that are tractable and can be implemented computationally. In the early days, top-down models have been mostly descriptive as they operated at the level of conceptual entities (e.g., collections of objects present in a scene, to be evaluated in terms of a mental model of which objects might be more important to a particular task of interest (Ballard, Hayhoe, & Pelz, 1995)), and hence have lacked generalized computational implementations (because, for example, algorithms that can robustly recognize objects in scenes were not available). As some of these hurdles have been alleviated by the advent of powerful new theories and algorithms for object recognition,

scene classification, decision making under uncertainty, machine learning, and formal description languages that can capture task descriptions and background knowledge, exciting new models are beginning to emerge which significantly surpass purely bottom-up approaches in their ability to predict attention behavior in complex situations. Implementing complete, autonomous computational models of top-down attention is particularly important to our theoretical understanding of task and context effects on attention, as these implementations remind us that some assumptions often made to develop abstract models may not give rise to tractable computational models (e.g., it is easy to note how humans are able to directly fixate a jar of jam when it is the next object required to make a sandwich (Land & Hayhoe, 2001) — but how did they know that a jar is present and where it is located, if not through some previous bottom-up analysis of the visual environment?). As we further discuss in this chapter, these new computational models also often blur the somewhat artificial dichotomy between bottom-up and top-down processing, since the so-called top-down models do rely to a large extent onto a bottom-up flow of incoming information that is merged with goals and task demands to give rise to the decision of where to attend next.

To frame the concepts exposed so far into a broader picture, we refer to Figure 2.1 as a possible anchor to help organize our thoughts and discussions of how different elements of a visual scene understanding system may work together in the primate brain. In practice, few system-level efforts have included all components mentioned in Figure 2.1, and most of our discussion will focus on computational models that implement parts of such a system.

In what follows, we first examine in more details the key concepts of early bottom-up attention models (Section 2.3), and then provide an overview and comparison of many subsequent models that have provided new exciting insight into defining and computing bottom-up salience and attention (Section 2.4). We then turn to top-down models (Section 2.5), first motivating them from experimental evidence, and then examining in turn top-down models that modulate feature gains (Section 2.5.1), that derive spatial priors from the gist of the visual scene (Section 2.5.2), and that implement more complex information foraging and decision making schemes (Section 2.5.3). Finally, we discuss in Section 2.6 lessons learned from these models, including on the nature and interaction between bottom-up and top-down processes, and promising directions towards the creation of even more powerful combined bottom-up and top-down models.

2.3 Early bottom-up attention concepts and models

Early attention models have been primarily influenced by the Feature Integration Theory (A. Treisman & Gelade, 1980), according to which incoming visual information is first analyzed by early visual neurons which are sensitive to elementary visual features of the stimulus (e.g., colors, orientations, etc). This analysis, operated in parallel over the entire visual field and at multiple spatial and temporal scales, gives rise to a number of cortical feature maps, where each map represents the amount of a given visual feature at any location in the visual field. Attention is then the process by which a region in space is selected and features within that region are re-assembled or bound back together to yield more complex object representations (Figure 2.2.a). Koch and Ullman (Koch & Ullman, 1985) extended the theory by advancing the concept of a single topographic and scalar *saliency map*, receiving inputs from the feature maps, as a computationally efficient representation upon which to operate the selection of where to attend next: A simple maximum-detector or *winner-take-all* neural network (Arbib & Didday, 1971) was proposed which would simply pick the next most salient location as the next attended one, while an active *inhibition-of-return* (M. Posner, 1980) mechanism would later inhibit that location and thereby allow attention to shift as the winner-take-all network would pick the next most salient location (Figure 2.2.b). From these ideas, a number of fully computational models started to be developed (e.g., Figure 2.2.c,d).

At the core of these early models is the notion of visual salience, a signal that is computed in a stimulus-driven manner and which indicates that some location is significantly different from its surroundings and is worthy of attention. Early models computed visual salience from bottom-up features in several feature maps, including luminance contrast, red-green and blue-yellow color opponency, and oriented edges (Itti, Koch, & Niebur, 1998). While visual salience is sometimes carelessly described as a physical property of a visual stimulus, it is important to remember that salience is the consequence of an interaction of a stimulus with other stimuli, as well as with a visual system (biological or artificial). For example, a color-blind person will have a dramatically different experience of visual salience than a person with normal color vision, even when

both look at exactly the same colorful physical scene. Nevertheless, because visual salience is believed to primarily arise from fairly low-level and stereotypical computations in the early stages of visual processing, the factors contributing to salience are generally quite comparable from one observer to the next, leading to similar experiences across a range of observers and of viewing conditions.

The essence of salience lies in enhancing the neural and perceptual representation of locations whose local visual statistics significantly differ from the broad surrounding image statistics, in some behaviorally relevant manner. This basic principle is intuitively motivated as follows. Imagine a simple search array as depicted in Figure 2.2.e, where one bar pops-out because of its unique orientation. Now imagine examining a feature map which is tuned to stimulus intensity (luminance) contrast: because there are many white bars on a black background, early visual neurons sensitive to local intensity contrast will respond vigorously to each of the bars (distractors and target alike, since all have identical intensity). Based on the pattern of activity in this map, in which essentially every bar elicits a strong peak of activity, one would be hard pressed to pick one location as being clearly more interesting and worthy of attention than any of the others. Intuitively, hence, one might want to apply some normalization operator N(.) which would give a very low overall weight to this map's contribution to the final saliency map. The situation is quite different when examining a feature map where neurons are tuned to local vertical edges. In this map, one location (where the single roughly vertical bar is) would strongly excite the neural feature detectors, while all other locations would elicit much weaker responses. Hence, one location clearly stands out and hence becomes an obvious target for attention. It would be desirable in this situation that the normalization operator N(.) give a high weight to this map's contribution to the final saliency map (Itti, Koch, & Niebur, 1998; Itti & Koch, 2000a, 2001b).

Early bottom-up attention models have created substantial interest and excitement in the community, especially as they were shown to be applicable to an unconstrained variety of stimuli, as opposed to more traditional computer vision approaches at the time, which often had been designed to solve a specific task in a specific environment (e.g., detect human faces in photographs taken from a standing human viewpoint (Viola & Jones, 2001)). Indeed, no parameter tuning nor any prior knowledge related to the contents of the images or video clips to be processed was necessary for any of many early results, as the exact same model processed psychophysical stimuli, filmed outdoors scenes, Hollywood movie footage, video games, and robotic imagery. This gave rise to model prediction results that included, for example, the reproduction by Itti *et al.*'s model of human behavior in visual search tasks (e.g., pop-out versus conjunctive search (Itti & Koch, 2000a)); demonstration of strong robustness to image noise (Itti, Koch, & Niebur, 1998); automatic detection of traffic signs and other salient objects in natural environments filmed by a consumer-grade color video camera (Itti & Koch, 2001d); the detection of pedestrians in natural scenes (Miau, Papageorgiou, & Itti, 2001); and of military vehicles in overhead imagery (Itti, Gold, & Koch, 2001); and — most importantly — the widely demonstrated ability of the model to predict where humans look when freely images or videos that range from search arrays to fractals to satellite images to everyday indoors and outdoors scenes (Parkhurst et al., 2002; Peters, Iyer, Itti, & Koch, 2005).

2.4 Flourishing of bottom-up models

Following initial success, many research groups started exploring the notions of bottom-up attention and visual salience, which has given rise to many new computational models. We summarize 53 bottom-up models along 13 different factors in Figure 2.3. A thorough examination of all models is certainly not feasible in this limited space. Instead, we highlight below the main trends in seven categories that span from strong inspiration from biological vision to more abstract mathematical definitions and implementations of the concept of saliency. Models can coarsely be categorized as follows (but also see (Tsotsos & Rothenstein, 2011) for another possible taxonomy). Please note that some models fall under more than one category.

Cognitive models. Research into saliency modeling escalated after Itti *et al.*'s (Itti, Koch, & Niebur, 1998) implementation of Koch and Ullman's (Koch & Ullman, 1985) computational architecture based on the Feature Integration Theory (A. Treisman & Gelade, 1980). In cognitive models, which were the first ones to approach the problem of algorithmically computing saliency in arbitrary digital images, an input image is decomposed into a set of feature maps across spatial scales which are then linearly or non-linearly normalized and combined to form a master saliency map. An important element of this theory is the idea of center-surround which defines saliency as distinctiveness of an image region to its immediate surroundings. Almost

all saliency models are directly or indirectly inspired by cognitive concepts of visual attention (e.g., (Le Meur et al., 2006; Marat et al., 2009)).

Information-theoretic models. Stepping back from biological implementation machinery, models in this category are based on the premise that localized saliency computations serve to maximize information sampled from one's environment. These models assign higher saliency values to scene regions with rare features. Information of visual feature F is $I(F) = -log\ p(F)$ which is inversely proportional to the likelihood of observing F (i.e., $p(F)$). By fitting a distribution $P(F)$ to features (e.g., using Gaussian Mixture Model or Kernels), rare features can be immediately found by computing $P(F)^{-1}$ in an image. While in theory using any feature space is feasible, usually these models (inspired by efficient coding representations in visual cortex) utilize a sparse set of basis functions (using ICA filters) learned from a repository of natural scenes. Some basic approaches in this domain are AIM (Bruce & Tsotsos, 2005), Rarity (MANCAS, 2007), LG (Local + Global image patch rarity) (Borji & Itti, 2012), and incremental coding length models (Hou & Zhang, 2008).

Graphical models. Graphical models are generalized Bayesian models which have been employed for modeling complex attention mechanisms over space and time. Torralba (Torralba, 2003) proposed a Bayesian approach for modeling contextual effects on visual search which was later adopted in the SUN model (Zhang, Tong, Marks, Shan, & Cottrell, 2008) for fixation prediction in free viewing. Itti and Baldi (Itti & Baldi, 2005) defined surprising stimuli as those which significantly change beliefs of an observer. Harel *et al.* (GBVS) (Harel, Koch, & Perona, 2007) propagated similarity of features in a fully connected graph to build a saliency map. Avraham (Avraham & Lindenbaum, 2010), Jia Li *et al.*, (L. Li & Fei-Fei, 2010), and Tavakoli *et al.* (Rezazadegan Tavakoli, Rahtu, & Heikkilä, 2011), have also exploited Bayesian concepts for saliency modeling.

Decision theoretic models. This interpretation states that attention is driven optimally with respect to the end task. Gao and Vasconcelos (Gao & Vasconcelos, 2004) argued that for recognition, salient features are those that best distinguish a class of objects of interest from all other classes. Given some set of features $X = \{X_1, \cdots, X_d\}$, at locations l, where each location is assigned a class label Y with $Y_l = 0$ corresponding to background and $Y_l = 1$ indicates objects of interest, saliency is then a measure of mutual information (usually Kullback-Leibler divergence (KL)), computed as $I(X, Y) = \sum_{i=1}^{d} I(X_i, Y)$. Besides having good accuracy in predicting eye fixations, these models have been very successful in computer vision applications (e.g., anomaly detection and object tracking).

Spectral-analysis models. Instead of processing an image in the spatial domain, these models derive saliency in the frequency domain. This way, there is no need for image processing operations such as center-surround or segmentation. Hou and Zhang (Hou & Zhang, 2007) derive saliency for an image with amplitude $\mathcal{A}(f)$ and phase $\mathcal{P}(f)$ as follows. The log spectrum $\mathcal{L}(f)$ is computed from the down-sampled image. From $\mathcal{L}(f)$, the spectral residual $\mathcal{R}(f)$ is obtained by multiplying $\mathcal{L}(f)$ with $h_n(f)$ which is an $n \times n$ local average filter and subtracting the result from itself. Saliency map is then the inverse Fourier transform of the exponential of amplitude plus phase (i.e., $\mathcal{S}(x) = \mathcal{F}^{-1}\big[exp\big(\mathcal{R}(f) + \mathcal{P}(f)\big)\big]$). The saliency of each point is squared to indicate the estimation error and is then smoothed with a Gaussian filter for better visual effect. Bian and Zhang (Bian & Zhang, 2009) and Guo *et al.* (Guo & Zhang, 2010) proposed spatio-temporal models in the spectral domain.

Pattern classification models. Models in this category use machine learning techniques to learn "stimuli-saliency" mappings from image features to eye fixations. They estimate saliency s; $p(s|f)$ where f is a feature vector which could be the contrast of a location and its surrounding neighborhood. Kienzle *et al.* (Kienzle, Wichmann, Schölkopf, & Franz, 2007), Peters and Itti (Peters & Itti, 2007), and Judd *et al.* (Judd et al., 2009a) used image patches, scene gist, and a vector of several features at each pixel, respectively and used classical SVM and Regression classifiers for learning saliency. In an extension of Judd model, Borji (Borji, 2012) showed that using a richer set of features including bottom-up saliency maps of other models and within-object regions (e.g., eye within faces) along with a boosting classifier leads to higher fixation predicting accuracy. Tavakoli *et al.* (Rezazadegan Tavakoli et al., 2011), used sparse sampling and kernel density estimation to estimate the above probability in a Bayesian framework. Note that some of these models may not be purely bottom-up since they use features that guide top-down attention, for example faces or text (Judd et al., 2009a; Cerf, Harel, Einhäuser, & Koch, 2008).

Other models. Some other models exist that do not easily fit into our categorization. For example, Seo and Milanfar (Seo & Milanfar, 2009) proposed self-resemblance of local image structure for saliency detection.

The idea of decorrelation of neural response was used for a normalization scheme in the Adaptive Whitening Saliency (AWS) model (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2009). Kootstra *et al.* (Kootstra, Nederveen, & De Boer, 2008) developed symmetry operators for measuring saliency and Goferman *et al.* (Goferman, Zelnik-Manor, & Tal, 2010) proposed a context-aware saliency detection model with successful applications in re-targeting and summarization.

An important trend to consider is that over the past years starting from (Liu, Sun, Zheng, Tang, & Shum, 2007), models have begun to diverge into two different classes: models of *fixation prediction* and models of *salient region detection*. While the goal of the former models is to predict locations that grab attention, the latter models attempt to segment the most salient object or region in a scene. A saliency operator is usually used to estimate the extent of the object that is predicted to be the most likely first attended object. Evaluation is often done by measuring precision-recall of saliency maps of a model against ground-truth data (explicit saliency judgments of subjects by annotating salient objects or clicking on locations). Some models in two categories have compared themselves against each other, without being aware of the distinction.

Figure 2.3 shows a list of models and their properties according to thirteen qualitative criteria derived from behavioral and computational studies. The majority (53 out of 65) of covered attention models consists of bottom-up models, indicating that at least from a computational perspective it is easier to formulate attention guidance mechanisms based on low-level image features. This is reinforced by the existence of several established benchmark datasets and standard evaluation scores for bottom-up models. The situation is the opposite for top-down attention modeling although we have recently initiated an effort to share data and code (Borji, Sihite, & Itti, 2012a, 2012c).

A brief comparison of saliency maps of 26 models on a few test images (Figure 2.4) shows large differences in appearance of the maps generated by different models. Some models generate very sparse maps while others are smoother. This makes fair model comparison a challenge since some scores may be influenced by smoothness of a map (Tatler et al., 2005). Recently, Borji *et al.* (Borji et al., 2012b) performed a detailed investigation of models to quantify their correlations with human attentional behavior. This study suggests that so far the so-called "shuffled AUC (Area Under the ROC Curve)" score (Zhang et al., 2008) is the most robust (this score uses distributions of human fixations on other stimuli than the one being scored to establish a baseline, which attenuates the effects of certain biases in eye movements datasets, the strongest being a bias towards looking preferentially near the center of any image). Results are shown in Figure 2.5. This model evaluation shows a gap between current models and human performance. This gap is smaller for some datasets, but overall exists. Discovering and adding more top-down features to models will hopefully boost their performance. The analysis also shows that some models are very effective (e.g., HouNIPS, Bian, HouCVPR, Torralba, and Itti-CIO2 in Figure 2.5) and also very fast providing a trade-off between accuracy and speed necessary for many applications.

Despite past progress in bottom-up saliency modeling and fixation prediction while freely viewing natural scenes, several open questions remain that should be answered in the future. The most confusing one is that of "center bias," whereby humans often appear to preferentially look near an image's center. It is believed to be largely caused by stimulus bias (e.g., photographer bias, whereby photographers tend to frame interesting objects near the image center). Collecting fixation datasets with no or less center bias, and studying its role on model evaluation needs to be addressed with natural scenes (see, e.g., (Parkhurst et al., 2002; Peters et al., 2005) for unbiased artificial datasets of fractal images). As opposed to saliency modeling on static scenes, the domain of spatio-temporal attention remains less explored (See (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; H. Wang, Freeman, Merriam, Hasson, & Heeger, 2012) for examples). Emphasis should be on finding cognitive factors (e.g., actor, non-actor) rather than simple bottom-up features (e.g., motion, flicker, or focus of expansion), and some of the top-down models discussed below have started to explore how more semantic scene analysis can influence attention. Another aspect is the study of attention on affective and emotional stimuli. Although a database of fixations on emotional images has been gathered by Ramanathan *et al.* (Ramanathan et al., 2010), it is still not clear whether current models can be extended to explain such fixations.

2.5 Top-down guidance of attention by task demands

Research towards understanding the mechanisms of top-down attention has given rise to two broad classes of models: models which operate on semantic content, and models which operate on raw pixels and images. Models in the first category are not fully computational in the sense used in the present chapter, in that they require that an external expert (typically, one or more humans) first pre-processes raw experimental recordings, often to create semantic annotations (e.g., translate from recorded video frames and gaze positions into sequences that describe which objects were being looked at). For example, in a block copying task (Ballard et al., 1995), the observers' algorithm for completing the task was revealed by their pattern of eye movements: first select a target block in the model by fixating it, then find a matching block in the resource pool, then revisit the model to verify the block's position, then fixate the workspace to place the new block in the corresponding position. Other studies have used naturalistic interactive or immersive environments to give high-level accounts of gaze behavior in terms of objects, agents, "gist" of the scene, and short-term memory (A. Yarbus, 1967; J. Henderson & Hollingworth, 1999; R. Rensink, 2000; Land & Hayhoe, 2001; Sodhi et al., 2002a; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003), to describe, for example, how task-relevant information guides eye movements while subjects make a sandwich (Land & Hayhoe, 2001; Hayhoe et al., 2003) or how distractions such as setting the radio or answering a phone affect eye movements while driving (Sodhi et al., 2002a).

While such perceptual studies have provided important constraints regarding goal-oriented high-level vision, additional work is needed to translate these descriptive results into fully-automated computational models that can be used in the application domains mentioned above. That is, although the block copying task reveals observers' algorithm for completing the task, it does so only in the high-level language of "workspace" and "blocks" and "matching." In order for a machine vision system to replicate human observers' ability to understand, locate, and exploit such visual concepts, we need a "compiler" to translate such high-level language into the assembly language of vision—that is, low-level computations on a time-varying array of raw pixels. Unfortunately, a general computational solution to this task is tantamount to solving computer vision.

From behavioral and in particular eye-tracking experiments during execution of real-world tasks, several key computational factors can be identified which can be implemented in computational models (Figure 2.6):

- **Spatial biases**, whereby a given high-level task or top-down set may make some region of space more likely to contain relevant information. For example, when the task is to drive, it is important to keep our eyes on the road (Figure 2.6.a). We describe below how bottom-up attention models can be enhanced by considering such task-driven spatial constraints, for example to suppress salient stimuli that lie outside the task-relevant region of visual space. These models are motivated by both psychophysical and physiological evidence of spatial biasing of attention based on both short-term and long-term top-down cues (Chun & Jiang, 1998; Summerfield, Rao, Garside, & Nobre, 2011), resulting in enhancement of attended visual regions and suppression of the un-attended ones (Brefczynski & DeYoe, 1999; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999);

- **Feature biases**, whereby the task may dictate that some visual features (e.g., some colors) are more likely associated with items of interest than other features (Figure 2.6.b). Bottom-up models can also be enhanced to account for feature biases, for example by modulating according to top-down goals the relative weights by which different feature maps contribute to a saliency map (e.g., when searching for a blue item, increase the gain of blue-selective feature maps). These models also are motivated by experimental studies of so-called feature-based attention (Treue & Martinez Trujillo, 1999; Saenz, Buracas, Boynton, et al., 2002; Zhou & Desimone, 2011; Martinez-Trujillo, 2011) and, in particular, recent theories and experiments investigating the role of the pulvinar nucleus in carrying out such biases (Baluch & Itti, 2011; Saalmann, Pinsk, Wang, Li, & Kastner, 2012);

- **Object-based and cognitive biases**, whereby knowing about objects, about how they may interact with each other, and about how they obey the laws of physics such as gravity and friction, may help humans make more efficient decisions of where to attend next to achieve a certain top-down goal (e.g., playing cricket, Figure 2.6.c, or making a sandwich, Figure 2.6.d). Models can be taught how to recognize objects and possibly other aspects of the world, to enable semantic reasoning that may

give rise to these more complex top-down attention behaviors. These models are also motivated by recent experimental findings (Võ & Henderson, 2009; Schmidt & Zelinsky, 2009; A. Hwang, Wang, & Pomplun, 2011).

We review top-down models that have implemented these strategies below. Although spatial biases have historically been studied first, we start with feature biasing models as those are conceptually simpler extensions to the bottom-up models described in the previous sections.

2.5.1 Top-down biasing of bottom-up feature gains

A simple strategy to include top-down influences in a computational attention model is to modulate the low-levels of visual processing of the model according to the top-down task demands. This embodies the concept of *feature-based attention*, whereby increased neural response can be detected in monkeys and humans to visual locations which contain features that match a feature of current behavioral interest (e.g., locations that contain upward moving dots when the animal's task is to monitor upward motion (Treue & Martinez Trujillo, 1999; Saenz et al., 2002; Zhou & Desimone, 2011).

While the idea of top-down feature biases was already present in early conceptual models like the *Guided Search* theory (J. Wolfe, 1994) and *FeatureGate* (Cave, 1999), the question for computational modelers has been how exactly the feature gains should be adjusted to yield optimal expected enhancement of a desired target among unwanted distractors (Figure 2.7). Earlier models have used supervised learning techniques to compute feature gains from example images where targets of interest had been manually indicated (Itti & Koch, 2001d; Frintrop, Backer, & Rome, 2005; Borji, Ahmadabadi, & Araabi, 2011). A more recent approach uses eye movement recordings to determine these weights (Zhao & Koch, 2011). Interestingly, it has recently been proposed that an optimal set of weight can be computed in closed-form given distributions of expected features for both targets of interest to the task and irrelevant clutter or distractors. In this approach, each feature map is characterized by a target-to-distractor response ratio (or signal-to-noise ratio, SNR), and feature maps are simply assigned a weight that is inversely proportional to their SNR (Navalpakkam & Itti, 2006, 2007). In addition to giving rise to a fully computational model, this theory has also been found to explain many aspects of human guided search behavior (Navalpakkam & Itti, 2007; Serences & Saproo, 2010)

Note how the models of Figure 2.7 start introducing a blur between the notions of bottom-up and top-down processing. In these models, indeed, the effect of top-down knowledge is to modulate the way in which bottom-up computations are carried out. This opens the question of whether a pure bottom-up state may ever exist, and what the corresponding gain values may be (e.g., unity as has been assumed in many models?). These questions are also being raised by a number of recent experiments (Theeuwes, 2010a; Awh, Belopolsky, & Theeuwes, 2012), and are further discussed below.

2.5.2 Spatial priors and scene context

In a recent model, Ehinger *et al.* (Ehinger et al., 2009) investigated whether a model that, in addition to a bottom-up saliency map, learns spatial priors about where people may appear and a feature prior about what people may look like, would better predict gaze patterns of humans searching for people. Indeed, several earlier studies had suggested that bottom-up models, while widely demonstrated to correlate with human fixations during free viewing, may not well predict fixations of participants once they are given a top-down task, for example a search task (Zelinsky, Zhang, Yu, Chen, & Samaras, 2006; Foulsham & Underwood, 2007; J. Henderson, Brockmole, Castelhano, & Mack, 2007; Einhäuser, Rutishauser, & Koch, 2008). One may argue, however, especially in the light of our above discussion of whether pure bottom-up salience is a valid concept, that in these experiments the models were at an unfair disadvantage: Human participants had been provided with some information which had not been communicated to models (e.g., search for a specific target, shown for 1 second before the search (Zelinsky et al., 2006); search of objects in a category or for a specific object (Foulsham & Underwood, 2007); search for the small bullseye pattern or for a local higher-contrast region (Einhäuser, Rutishauser, & Koch, 2008); or count people (J. Henderson et al., 2007)). Ehinger *et al.* addressed this by proposing a model that combines three sources of information (Figure 2.8.a): First, a "scene context" map was derived from learning the associations between holistic or global scene features (coarsely capturing the gist of the scene (Torralba, 2003)) and the locations where humans appeared in scenes with given holistic features (trained over 1880 example images). This map, which

is of central interest to this section of our chapter, thus learned the typical locations where humans were expected to appear in different views of street scenes. This learning step produces a prior on locations that can be used to filter out salient responses in locations that are highly unlikely to contain the target (e.g., in the sky, assuming that no human was seen flying in the training dataset). Second, a person detector was run in a sliding window manner over the entire image, creating a "target features" map that highlighted locations that closely look like humans. This provides an alternative to learning feature gains as discussed above; instead, an object detector algorithm is trained for the desired type of target. While possibly more efficient than gain modulation, this approach suffers from lower biological plausibility. (See (R. P. N. Rao, Zelinsky, Hayhoe, & Ballard, 2002; Orabona et al., 2005) for related models). Third and finally, a standard bottom-up "saliency map" provided additional candidate locations (also see (Oliva, Torralba, Castelhano, & Henderson, 2003) for earlier related work, integrating only saliency and scene context). Ehinger *et al.* found that the model which combined all three maps outperformed any of the three component models taken alone (Figure 2.8.b).

In a related model, Peters & Itti (Peters & Itti, 2007) also used a combination of bottom-up saliency maps and top-down spatial maps derived from the holistic gist of the scene, but their top-down maps were directly learned from eye movements of human observers, playing the same 3D video games as would be used for testing (games included driving, exploration, flight combat, etc.; note that since players control the game's virtual camera viewpoint, each run of such game gives rise to a unique set of viewpoints and of generated scenes). The bottom-up component of this model is based on the Itti-Koch saliency model (Itti, Koch, & Niebur, 1998), which predicts interesting locations based on low-level visual features such as luminance contrast, color contrast, orientation, and motion. The top-down component is based on the idea of "gist," which in psychophysical terms is the ability of people to roughly describe the type and overall layout of an image after only a very brief presentation (F. Li, VanRullen, Koch, & Perona, 2002), and to use this information to guide subsequent target searches (Torralba, 2003). This model (Figure 2.8.c) decomposes each video frame into a low-level image signature intended to capture some of the properties of "gist" (Siagian & Itti, 2007), and learns to pair the low-level signatures from a series of video clips with the corresponding eye positions; once trained, it generates predicted gaze density maps from the gist signatures of previously unseen video frames. To test these bottom-up and top-down components, we compared their predicted gaze density maps with the actual eye positions recorded while people interactively played video games (Figure 2.8.d).

2.5.3 More complex top-down models

Many top-down models have been proposed which include higher degrees of cognitive scene understanding. Already in the late 1990's several models included a top-down component that decided where to look next based on what had been observed so far (e.g., (Rybak, Gusakova, Golovan, Podladchikova, & Shevtsova, 1998; Schill, Umkehrer, Beinlich, Krieger, & Zetzsche, 2001); also see (Itti & Koch, 2001b) for review). In robotics, the notion of combining or alternating between different behaviors (such as exploration versus search, or bottom-up versus top-down) has also led to several successful models (Sprague & Ballard, 2003; Forssén et al., 2008; Burattini, Rossi, Finzi, & Staffa, 2010; T. Xu, Kuhnlenz, & Buss, 2010). More recently, and our focus here, probabilistic inference and reasoning techniques, very popular in computer vision, have started to be used in attention models.

In many recent models, the saliency map of bottom-up models is conserved as a data-driven source of information for an overarching top-down system (more complicated than the feature or spatial biasing described above). For example, Boccignone & Ferraro (Boccignone & Ferraro, 2004) develop an overt attention system where the top-down component is a random walker that follows an information foraging strategy over a bottom-up saliency map. They demonstrate simulated gaze patterns that better match human distributions (Tatler, Hayhoe, Land, & Ballard, 2011). Interesting related models have been proposed where bottom-up and top-down attention interact through object recognition (Ban, Kim, & Lee, 2010; S. Lee et al., 2011), or by formulating a task as a classification problem with missing features, with top-down attention then providing a choice process over the missing features (Hansen, Karadogan, & Marchegiani, 2011).

Of growing recent interest is the use of probabilistic reasoning and graphical models to explore how several sources of bottom-up and top-down information may combine in a Bayesian-optimal manner. For example, the model of Akamine *et al.* (Akamine et al., 2012) (also see (Kimura, Pang, Takeuchi, Yamato,

& Kashino, 2008)), which employs probabilistic graphical modeling techniques and considers the following factors, interacting in a dynamic Bayesian network (Figure 2.9.a): On the one hand, input video frames give rise to deterministic saliency maps. These are converted into stochastic saliency maps via a random process that affects the shape of salient blobs over time (e.g., dynamic Markov random field (Kimura et al., 2008)). An eye focusing map is then created which highlights maxima in the stochastic saliency map, additionally integrating top-down influences from an eye movement pattern (a stochastic selection between passive and active state with a learned transition probability matrix). The authors use a particle filter with Markov chain Monte-Carlo (MCMC) sampling to estimate the parameters; this technique often used in machine learning allows for fast and efficient estimation of unknown probability density functions. Although the top-down component is quite simple in this version of the model, it is easy to see how more sophisticated top-down and contextual influences could be integrated into the dynamic Bayesian network framework of Kimura *et al.* Several additional recent related models using graphical models have been proposed (e.g., (Chikkerur, Serre, Tan, & Poggio, 2010)).

Although few have been implemented as fully computational models, several efforts have started to develop models that perform reasoning over objects or other scene elements to make a cognitive decision of where to look next (Navalpakkam & Itti, 2005; Yu, Mann, & Gosine, 2008; Beuter, Lohmann, Schmidt, & Kummert, 2009; Yu, Mann, & Gosine, 2012).

A a recent example, using probabilistic reasoning and inference tools, Borji *et al.* (Borji et al., 2012a) introduced a framework to model top-down overt visual attention based on reasoning, in a task-dependent manner, about objects present in the scene and about previous eye movements. They designed a Dynamic Bayesian Network (DBN) that infers probability distributions over attended objects and spatial locations directly from observed data. Two basic concepts in this model are 1) taking advantage of the sequence structure of tasks, which allows to predict the future fixations from past fixations and knowledge about objects present in the scene. Graphical models have indeed been very successful in the past to model sequences with applications in different domains, including biology, time series modeling, and video processing, and 2) computing attention at the object level. Since objects are essential building blocks in scenes, it is reasonable to assume that humans have instantaneous access to task-driven object-level variables (as opposed to only gist-like, scene-global, representations). Briefly, the model works by defining a Bayesian network over object variables that matter for the task. For example, in a video game where one runs a hot-dog stand and has to serve multiple hungry customers while managing the grill, those include raw sausages, cooked sausages, buns, ketchup, etc.(Figure 2.9.b). Then, existing objects in the scene, as well as the previous attended object, provide evidence toward the next attended object (Figure 2.9.b). The model also allows to read out which spatial location will be attended, thus allowing one to verify its accuracy against the next actual fixation of the human player. The parameters of the network are learned directly from training data in the same form as the test data (human players playing the game). This object-based model was significantly more predictive of eye fixations compared to simpler classifier-based models, also developed by the same authors, that map a signature of a scene to eye positions, several state-of-the-art bottom-up saliency models, as well as brute-force algorithms such as mean eye position (Figure 2.9.c). This points toward the efficacy of this class of models for modeling spatio-temporal visual data in presence of a task and hence a promising direction for future. Probabilistic inference in this model is performed over object-related functions which are fed from manual annotations of objects in video scenes or by state-of-the-art object detection models. (Also see (Y. Sun & Fisher, 2003; Y. Sun, Fisher, Wang, & Gomes, 2008) for models that consider objects, although they do not reason about object identities and task-dependent roles).

Finally, several computational models have started to explore making predictions that go beyond simply the next attended location. For example, Peters & Itti (Peters & Itti, 2008) developed a model that monitors in an online manner video frames and eye gaze of humans engaged in 3D video games, computing instantaneous measures of how well correlated the eye is with saliency predictions and with gist-based top-down predictions. They then learn to detect specific patterns in these instantaneous measures, which allows them to predict — up to several seconds in in advance — when players are about to fire a missile in a flight combat game, or to shift gears in a driving game. This model hence in essence estimates the intentions and predicts the future actions of the player. A related recent model was proposed by Doshi & Trivedi (Doshi & Trivedi, 2010) for active vehicle safety and driver monitoring. The system both computes bottom-up and top-down saliency maps from a video feed of the driver's view, and monitors the eye movements of the driver to better predict driver attention and gaze by estimating online the cognitive state and level of distraction of the driver.

Using similar principles and adding pattern classification techniques, Tseng *et al.* (Tseng et al., 2009) have recently introduced a model that uses machine learning to classify, from features collected at the point of gaze over a few minutes of television viewing, control subjects from patients with disorders that affect the attention and oculomotor systems. The model has been successfully applied to elderly subjects (classifying patients with Parkinson's disease vs. controls) as well as children (classifying children with Attention Deficit Hyperactivity Disorder vs. Fetal Alcohol Spectrum Disorder vs. controls well above chance). These recent efforts suggest that eye movement patterns in complex scenes do contain — like a drop of saliva — latent individual biomarkers, which the latest attention modeling and pattern classification techniques are now beginning to reliably decode.

2.6 Discussion and outlook

Our review shows that tremendous progress has been made in modeling both bottom-up and top-down aspects of attention computationally. Tens of new models have been developed, each bringing new insight into the question of what makes some stimuli more important to visual observers than other stimuli.

Our quantitative comparison of many existing computational models on three standard datasets (Figure 2.5) prompts at least two reactions: First, it is encouraging to see that several models perform significantly better than trivial models (e.g., a central Gaussian blob) or than older models (e.g., Itti-CIO2 model in Figure 2.5). Second, however, it is surprising that the ranking of model scores is quite substantially different from the chronological order in which models were published. Indeed, one would typically expect that for a new model to be recognized, it should demonstrate superior performance compared to the state of the art, and actually often this is the case — just using possibly different datasets, scoring metrics, etc. Thus one important conclusion of our study is that carrying out standardized evaluations is important to ensure that the field keeps moving forward (see (Borji, 2010) for a web-based effort in this direction).

Another important aspect of model evaluation is that currently almost all model comparisons and scoring are based on average performance over a dataset of images or video clips, where often the dataset has been hand-picked and may contain significant biases (Torralba & Efros, 2011). It may be more fruitful in the future to focus scoring on the most dramatic mistakes a model might make, or the worst-case disagreement between model and human observers. Indeed, average measures can easily be dominated by trivial cases if those happen often (e.g., we discussed earlier the notion of center bias and how a majority of saccades which humans make are aimed towards the centers of images), and models may be developed which perform well in these cases but miss conceptually important understanding of how attention may operate in the minority of non-trivial cases. In addition, departure from average performance measures may provide richer information about which aspects of attention are better captured by a given model (e.g., some models may perform better on some sub-categories or even instances of images than others).

As we described more models, and in particular started moving from bottom-up models to those which include top-down biases, the question arose of whether purely bottom-up models are indeed relevant to real life. In other words, is there such a state of human cognition where a default or unbiased form of salience may be computed and may guide gaze. Many experiments have assumed that free viewing, just telling observers to "watch and enjoy" stimuli presented to them, might be an acceptable approximation to this canonical unbiased state. However, it is trivially clear from introspection that cognition is not turned off during free viewing, and that what we look at in one instant triggers a range of memories, emotions, desires, cognitive inferences, etc. which all will ultimately influence where we look next. In this regard, it has been recently suggested that maybe only the initial volley of activity through visual cortex following stimulus onset may represent such canonical bottom-up saliency representation (Theeuwes, 2010a). If such is the case, then maybe comparing model predictions to sometimes rather long sequences of eye movements may be not be the best measure of how well a model captures this initial purely bottom-up attention.

Models where top-down influences serve to bias the bottom-up processing stages have also blurred the line between bottom-up and top-down. In fact, this is an important reminder that bottom-up and top-down influences are not mutually exclusive and do not sum to give rise to attention control (Awh et al., 2012). Instead, bottom-up and top-down often agree: The actor cognitively identified as the protagonist in a video clip may also move in such ways that he is the most salient. In fact, today's bottom-up may be nothing more than our former generations' top-down. Indeed, some bottom-up models have successfully integrated

high-level features such as human face detectors into their palette of feature maps (Judd et al., 2009a), which blurs again the line between bottom-up and top-down (these features are computed in a bottom-up manner from the image, but their very presence in a model is based on top-down knowledge that humans do strongly tend to look at faces in images (Cerf, Frady, & Koch, 2009)).

When there is a task, top-down influences on attention are often believed to dominate, though this remains controversial and depends both on the task and on the quantification method used (Zelinsky et al., 2006; Foulsham & Underwood, 2007; J. Henderson et al., 2007; Einhäuser, Rutishauser, & Koch, 2008; Greene, Liu, & Wolfe, 2012). In human vision, we should not forget the following: Purely top-down attention (i.e., making a purely volitional eye movement unrelated to any visual stimulus) is not a generally viable model, except maybe in blind persons. Others may make pure top-down eye movements from time to time, but certainly not always — that is, no matter how strongly one believes that top-down influences dominate, in the end controlling visual attention is a visually-guided behavior, and, as such, it is dependent on visual stimuli. This is important for future modeling efforts, as they attempt to tackle more complex tasks and situations, such as making a sandwich (Figure 2.6.d): A person may indeed look at the jar of jam because it is the next required object for the task (top-down guidance towards the jam). However, how did that person know where the jam is? In most cases, some bottom-up analysis (maybe in the past) must have provided that information (except maybe if the person was told where the jam is). Thus, modeling human behavior in complex tasks will likely require very careful control over the experimental setup, so that human participants are not given more information or additional priors that are not communicated to models (e.g., let a person look around before the task begins; see (Foulsham, 2012) for recent relevant data). This consideration echoes our earlier remark about making fair comparisons: If a model is not given the same information as a human participant (e.g., the model is not biased towards a search target or is not allowed to explore a scene before the task begins), likely the model will not perform as well, but we will also learn very little from such an experiment.

Another important challenge for models briefly mentioned above is dealing with sequence in eye movement data (i.e., scanpath) and with how to capture temporality in saccades. When comparing models, a model might be favored not only if it can predict exact saccade locations, but also their ordering and their individual times of occurrence. In free viewing, in spite of past efforts (Privitera & Stark, 2000), it is still not clear whether such sequential information is a strong factor of attention control and to what extent it depends on the subject or the asked question (A. Yarbus, 1967). Despite this, recently some researchers (e.g., (W. Wang et al., 2011)) have tried to develop models and scores to explain sequences of saccades. As opposed to free viewing, it seems that there is much more temporal information in saccades in presence of a task. For instance, assume an observer is viewing videos of two different tasks such as sandwich making or driving. It probably should not be very difficult to decode the task just from the sequential pattern of eye movements. This means that the task governs sequence of saccades when there is a task. In free-viewing, however, when subjects are asked to watch a static scene freely there might not be a unique instruction making them to saccade sequentially to certain places. Even if subjects are asked to watch the scene under different questions, chances are that sequence may not help to decode the task (Greene et al., 2012).

Our survey shows that the remaining gap between man and machine seems to a large extent to be in 3D+time scene understanding, which includes reconstruction of the 3D geometry of the scene, understanding temporal sequences of events, simulation and extrapolation of physics over time in that 3D environment (e.g., to extrapolate the trajectory of a ball as in Figure 2.6.c), and so on. This requires some degree of machine vision and scene understanding which is not yet solved in the general case. This means that future computational models of attention will need to bring to bear sophisticated machine vision algorithms for scene understanding, to provide the necessary parsing of visual inputs into tokens that can be reasoned upon and prioritized by attention.

Acknowledgements

Supported by the National Science Foundation (grant numbers BCS-0827764 and CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and 62221-NS), the U.S. Army (W81XWH-10-2-0076), and Google. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

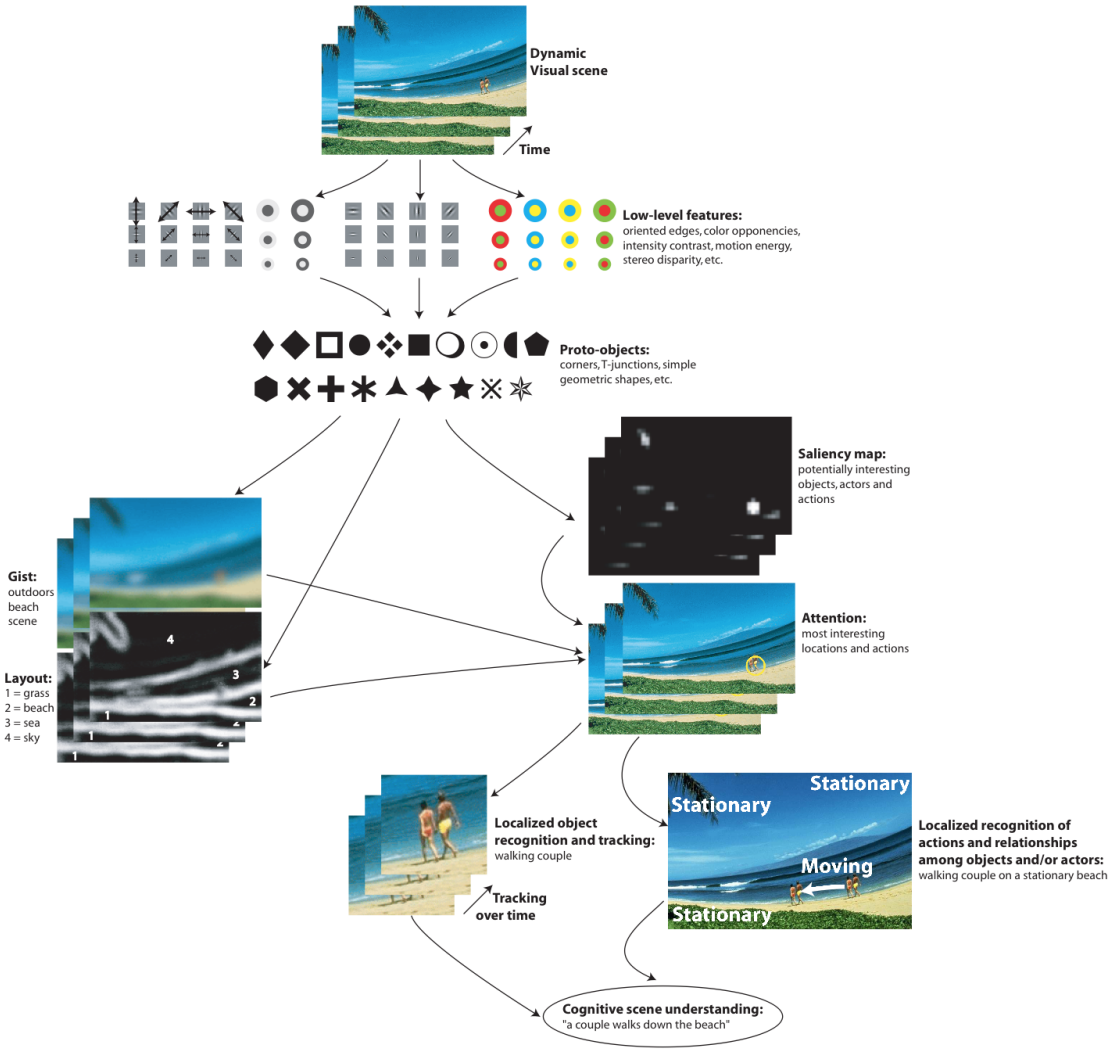


Figure 2.1: Minimal attention-based architecture for complex dynamic visual scene understanding. This diagram augments the triadic architecture of Rensink (2000), which identified three key components of visual processing: volatile (instantaneous) and parallel pre-attentive processing over the entire visual field, from the lowest-level features up to slightly more complex proto-object representations (top), identification of the setting (scene gist and layout; left), and attentional vision including detailed and more persistent object recognition within the spatially circumscribed focus of attention (right). Here we have extended Rensink's architecture to include a saliency map to guide attention bottom-up towards salient image locations, and action recognition in dynamic scenes. Also see Navalpakkam *et al.* (2005).

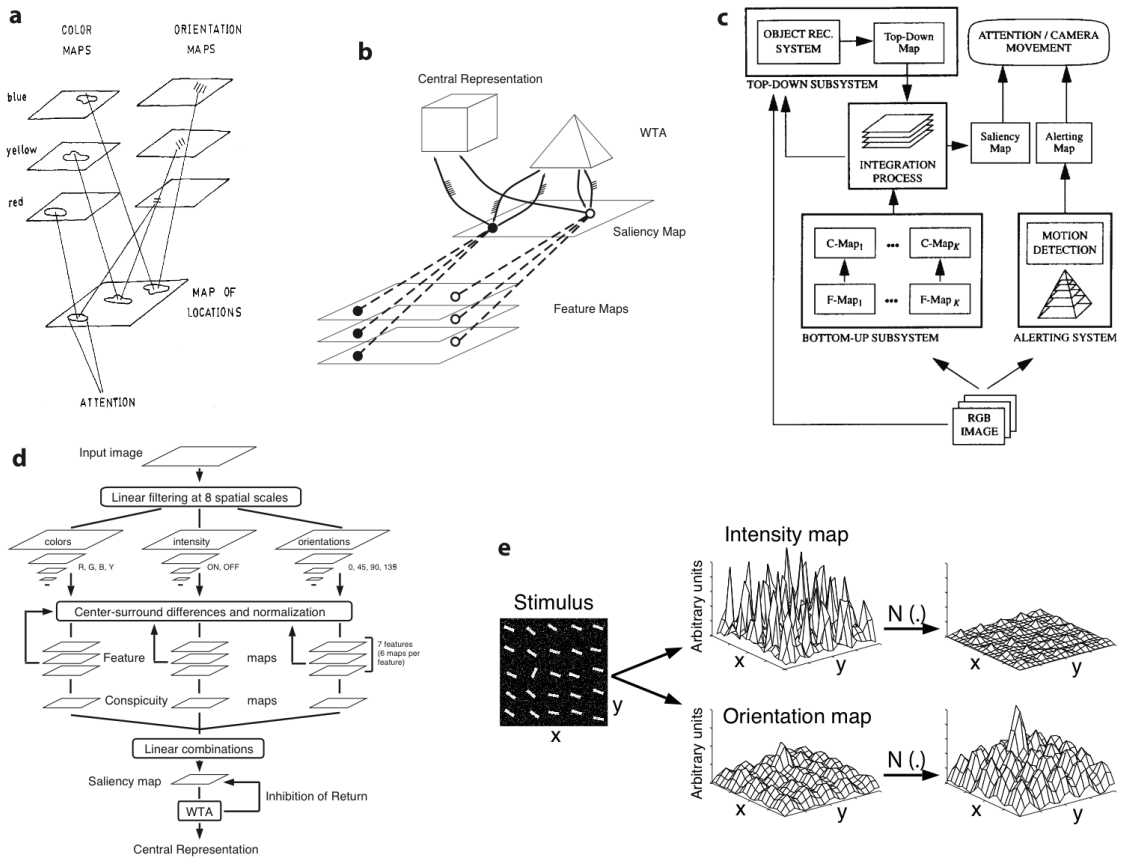


Figure 2.2: Early bottom-up attention theories and models. **(a)** Feature integration theory of Treisman & Gelade (1980) posits several feature maps, and a focus of attention that scans a map of locations and collects and binds features at the currently attended location. (from Treisman & Souther (1985)). **(b)** Koch & Ullman (1985) introduced the concept of a saliency map receiving bottom-up inputs from all feature maps, where a winner-take-all (WTA) network selects the most salient location for further processing. **(c)** Milanese *et al.* (1994) provided one of the earliest computational models. They included many elements of the Koch & Ullman framework, and added new components, such as an alerting subsystem (motion-based saliency map) and a top-down subsystem (which could modulate the saliency map based on memories of previously recognized objects). **(d)** Itti *et al.* (1998) proposed a complete computational implementation of a purely bottom-up and task-independent model based on Koch & Ullman's theory, including multiscale feature maps, saliency map, winner-take-all, and inhibition of return. **(e)** One of the key elements of Itti *et al.*'s model is to clearly define an attention interest operator, here denoted N(.), whereby the weight by which each feature map contributes to the final saliency map depends on how busy the feature map is. This embodies the idea that feature maps where one location significantly stands out from all others (as is the case in the orientation map shown) should strongly contribute to salience because they clearly vote for a particular location in space as the next focus of attention. In contrast, feature maps where many locations elicit comparable responses (e.g., intensity map shown) should not strongly contribute because they provide no clear indication of which location should be looked at next.

	No	Model	Year	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13
Bottom-up Visual Saliency Models for fixation prediction or salient region detection)	1	Itti et al.	1998	+	-	-	+	-	-	+	f	+	CIO	C	-	-
	2	Privitera & Stark	2000	+	-	-	+	-	-	+	f	+	-	O	-	Stark and Choi
	3	Salah et al.	2002	+	+	-	+	-	-	+	-	+	O	G	DR	Digit & Face
	4	Itti et al.	2003	+	-	+	+	+	+	+	f	+	CIOFM	C	-	-
	5	Torralba	2003	-	+	-	+	-	-	+	s	+	CI	B	DR	Torralba et al.
	6	Sun & Fisher	2003	+	-	-	+	-	-	+	-	-	CIO	G	-	-
	7	Gao & Vasconcelos	2004	-	+	-	+	-	-	+	s	-	DCT	D	DR	Brodatz, Caltech
	8	Ouerhani et al.	2004	+	-	-	+	-	-	+	f	+	CIO+Corner	C	CC	Ouerhani
	9	Boccignone & Ferraro	2004	+	-	+	-	+	-	+	f	-	Optical Flow	B	-	BEHAVE
	10	Frintrop	2005	+	+	+	+	+	+	+	f/s	+/-	CIOM	C	-	-
	11	Itti & Baldi	2005	+	-	+	+	+	+	-	f	+	CIOFM	B	KL, AUC	ORIG-MTV
	12	Ma et al.	2005	+	-	+	+	-	-	+	f	+	M*	O	-	-
	13	Bruce & Tsotsos	2006	+	-	-	+	-	+	+	f	+	DOG, ICA	I	KL, ROC	Bruce and Tsotsos
	14	Navalpakkam & Itti	2006	-	+	-	+	-	+	+	s	+	CIO	C	-	-
	15	Zhai & Shah	2006	+	-	+	+	+	-	+	f	+	SIFT	O	-	-
	16	Harel et al.	2006	+	-	-	+	-	-	+	f	+	IO	G	AUC	Bruce and Tsotsos
	17	Le Meur et al.	2006	+	-	-	+	-	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
	18	Walther & Koch	2006	+	-	-	+	-	+	+	f	+/-	CIO	C	-	-
	19	Peters & Itti	2007	+	+	+	+	+	-	+	i	+	CIOFM	P	KL, NSS	Peters and Itti
	20	Liu et al.	2007	+	-	-	+	-	-	+	f	-	Liu*	G	F-measure	Regional
	21	Shic & Scassellati	2007	+	-	+	+	+	-	+	f	+	CIOM	C	ROC	Shic and Scassellati
	22	Hou & Zhang	2007	+	-	-	+	-	+	+	f	+	FFT, DCT	S	NSS	DB of Hou and Zhang, 2007
	23	Cerf et al.	2007	+	+	-	+	-	+	+	f/s	+	CIO :)	C	AUC	Cerf et al.
	24	Le Meur et al.	2007	+	-	+	+	+	-	+	f	+	LM*	C	CC, KL	Le Meur et al.
	25	Mancas	2007	+	-	+	+	+	+	+	f	+	CI	I	CC	Le Meur et al.
	26	Guo et al.	2008	+	-	-	+	-	-	+	f	+	CIO	D	CC	Self data
	27	Zhang et al.	2008	+	-	-	+	-	+	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
	28	Hou & Zhang	2008	+	-	+	+	+	-	+	f	+	ICA	I	AUC, KL	Bruce and Tsotsos, ORIG
	29	Pang et al.	2008	+	+	+	+	+	-	+	f	+	CIOM	G	NSS	ORIG, Self data
	30	Kootstra et al.	2008	+	-	-	+	-	-	+	f	+	Symmetry	C	CC	Kootstra et al.
	31	Ban et al.	2008	+	-	+	+	+	-	+	f	+	CIO+SYM	I	-	-
	32	Rajashekar et al.	2008	+	-	-	+	-	-	+	f	+	R*	S	CC	Rajashekar et al.
	33	Aziz and Mertsching	2008	+	-	-	+	-	+	+	f	+	CO,size, sym	I	_	-
	34	Kienzle et al.	2009	+	-	-	+	-	-	+	f	+	I	P	K*	Kienzle et al.
	35	Marat et al.	2009	+	-	+	+	+	-	+	f	+	SM*	C	NSS	Marat et al.
	36	Judd et al.	2009	+	-	-	+	-	-	+	f	+	J*	P	AUC	Judd et al.
	37	Seo & Milanfar	2009	+	-	+	+	+	+	+	f	+	LSK	I	AUC, KL	Bruce and Tsotsos, ORIG
	38	Rosin	2009	+	-	-	+	-	-	+	f	+	C+ Edge	O	PR, F-measure	DB of Liu et al, 2007
	39	Yin Li et al.	2009	-	+	+	+	+	+	+	s	+	RGB	S	DR	DB of Hou and Zhang, 2007
	40	Bian & Zhang	2009	+	-	+	+	+	+	+	f	+	FFT	S	AUC	Bruce and Tsotsos
	41	Diaz et al.	2009	+	-	-	+	-	+	+	f	+	CIO	O	AUC	Bruce and Tsotsos
	42	Zhang et al.	2009	+	-	+	-	+	-	+	f	+	DOG, ICA	B	KL, AUC	Bruce and Tsotsos
	43	Achanta et al.	2009	+	-	-	+	-	-	+	f	+	DOG	S	PR	DB of Liu et al, 2007
	44	Gao et al.	2009	+	-	+	+	+	+	+	f	+	CIO	D	AUC	Bruce and Tsotsos
	45	Chikkerur et al.	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	B	AUC	Bruce and Tsotsos, Chikkerur
	46	Mahadaven & Vasconcelos	2010	+	-	+	-	+	-	+	-	+	I	D	DR, AUC	SVCL background data
	47	Avraham & Lindenbaum	2010	+	+	-	+	-	+	+	f/s	+/-	CIO	G	DR, CC	UWGT, Ouerhani et al.
	48	Jia Li et al.	2010	-	+	+	+	+	-	+	f	+	CIO	B	AUC	RSD, MTV, ORIG, Peters and Itti
	49	Guo et al.	2010	+	-	+	+	+	+	+	f/s	+/-	FFT	S	DR	Self data
	50	Borji et al.	2010	-	+	-	+	-	+	+	s	+/-	CIO	O	DR	-
	51	Goeferman et al.	2010	+	-	-	+	-	-	+	-	+	C :)	O	AUC	DB of Hou and Zhang, 2007
	52	Murray et al.	2011	+	-	-	+	-	-	+	f	+	CIO	C	AUC, KL	Bruce and Tsotsos, Judd et al.
	53	Wang et al.	2011	+	-	-	+	-	-	+	f	+	ICA	I	AUC	Self data
Top-down (general models)	54	McCallum	1995	-	+	-	+	-	+	-	i	+	-	R	-	Self data
	55	Rao et al.	1995	-	+	-	+	-	-	+	s	+	CIO	O	-	Self data
	56	Ramstrom & Christiansen	2002	-	+	-	+	-	-	+	-	+	CI	O	-	-
	57	Sprague & Ballard	2003	-	+	+	-	+	+	+	i	-	S*	R	-	-
	58	Renninger et al.	2004	-	+	-	+	-	+	-	s	-	Edgelet	I	DR	Self data
	59	Navalpakkam & Itti	2005	-	+	-	+	-	+	+	-	+	CIO	C	-	Self data
	60	Paletta et al.	2005	-	+	-	+	-	-	+	-	-	SIFT	R	DR	COIL-20, TSG-20
	61	Jodogne & Piater	2007	-	+	-	+	-	-	+	i	-	SIFT	R	-	-
	62	Butko & Movellan	2009	-	+	+	+	+	+	+	s	-	-	R	-	-
	63	Verma & McOwan	2009	+	-	-	+	-	+	-	s	-	CIO	O	-	-
	64	Borji et al.	2010	-	+	-	+	-	-	+	i	-	CIO	R	-	self data
	65	Borji et al.	2012	-	+	-	+	+	-	+	i	-	CIO	B	AUC, NSS	self data

Figure 2.3: Survey of bottom-up and top-down computational models, classified according to 13 factors. Factors in order are: Bottom-up (f_1), Top-down (f_2), Spatial (-)/Spatio-temporal (+) (f_3), Static (f_4), Dynamic (f_5), Synthetic (f_6) and Natural (f_7) stimuli, Task-type (f_8), Space-based(+)/Object-based(-) (f_9), Features (f_{10}), Model type (f_{11}), Measures (f_{12}), and Used dataset (f_{13}). In Task type (f_8) column: free-viewing (f); target search (s); interactive (i). In Features (f_{10}) column: CIO: color, intensity and orientation saliency; CIOFM: CIO plus flicker and motion saliency; M* = motion saliency, static saliency, camera motion, object (face) and aural saliency (Speech-music); LM* = contrast sensitivity, perceptual decomposition, visual masking and center-surround interactions; Liu* = center-surround histogram, multi-scale contrast and color spatial-distribution; R* = luminance, contrast, luminance-bandpass, contrast-bandpass; SM* = orientation and motion; J* = CIO, horizontal line, face, people detector, gist, etc; S* = color matching, depth and lines; :) = face. In Model type (f_{11}) column, R means that a model is based RL. In Measures (f_{12}) column: K* = used Wilcoxon-Mann-Whitney test (The probability that a random chosen target patch receives higher saliency than a randomly chosen negative one); DR means that models have used a measure of detection/classification rate to determine how successful was a model. PR stands for Precision-Recall. In dataset (f_{13}) column: Self data means that authors gathered their own data. For a detailed definition of these factors please refer to Borji & Itti (2012 PAMI).

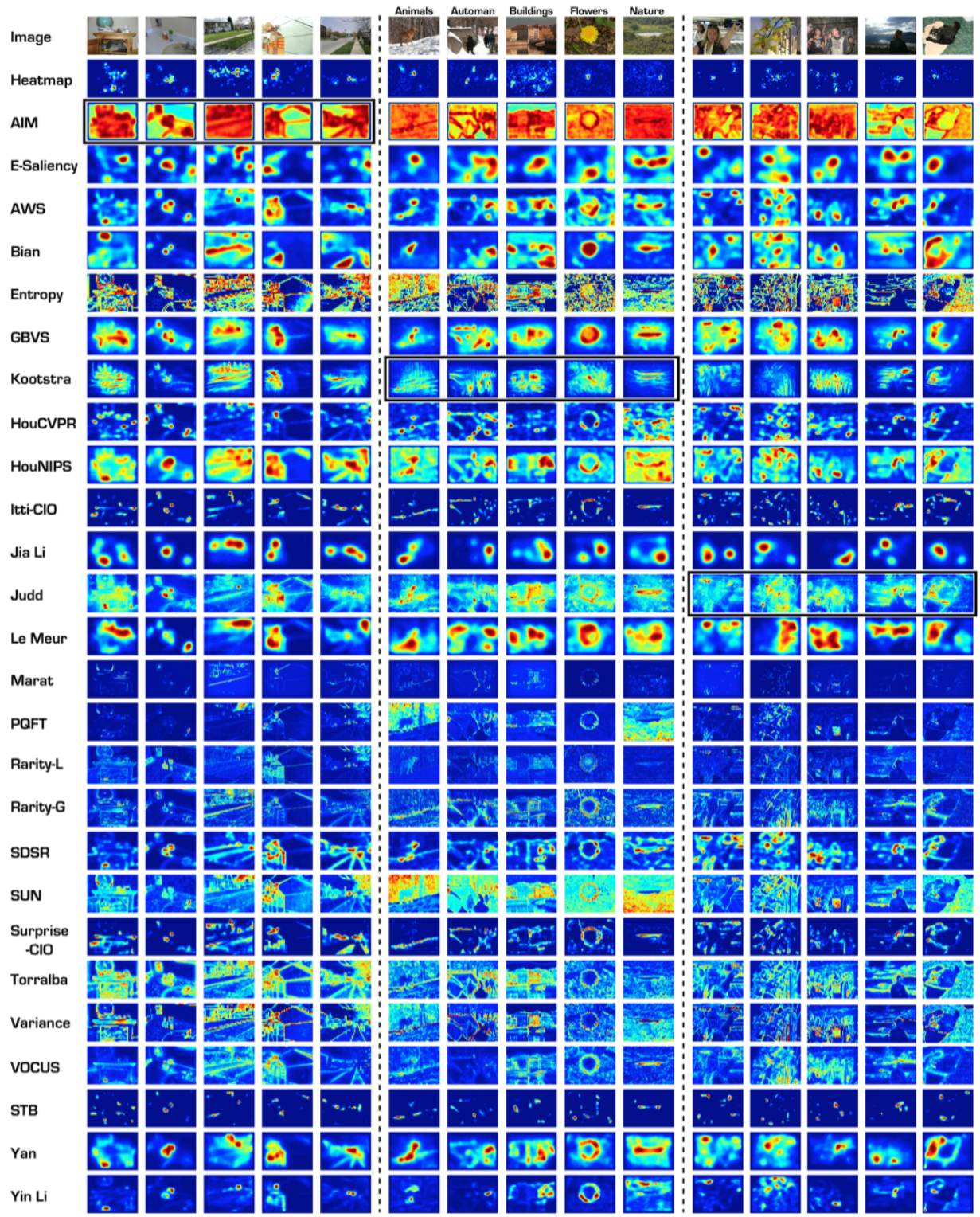


Figure 2.4: Example images (first row), human eye movement heatmaps (second row), and saliency maps from 26 computational models. The three vertical dashed lines separate the three datasets used (Bruce & Tsotsos, Kootstra & Schomacker, and Judd *et al.*). Black rectangles indicate the model originally associated with a given image dataset. Please see Borji *et al.* (2012 TIP) for additional details.

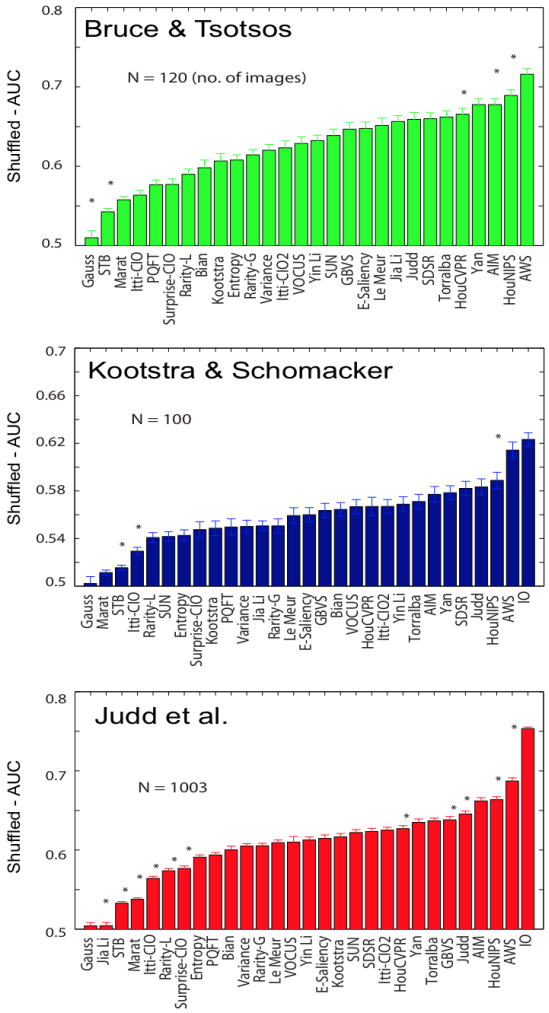


Figure 2.5: Ranking visual saliency models over three image datasets. Left column: Bruce & Tsotsos (2005), Middle column: Kootstra & Shomaker (2008), and Right column: Judd *et al.* (2009) using shuffled AUC score. Stars indicate statistical significance using t-test (95%, $p \leq 0.05$) between consecutive models. Error bars indicate standard error of the mean (SEM): $\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation and N is the number of images. The Judd model uses center feature, gist and horizon line, and object detectors for cars, faces, and human body. Itti-CIO2 is the approach proposed by Itti *et al.* (1998) that uses a normalization scheme known as Maxnorm: For each feature map, find the global max M and find the average m of all other local maxima. Then just weight the map by $(M - m)^2$. In the Itti-CIO method (Itti & Koch, 2000), normalization is: Convolve each map by a Difference of Gaussian(DoG) filter, cut off negative values, and iterate this process for a few times. As results show the Maxnorm normalization scheme performs better. In the literature, majority of models have been compared against Itti-CIO model. Please see Borji *et al.* (2012 TIP) for additional details on these results.

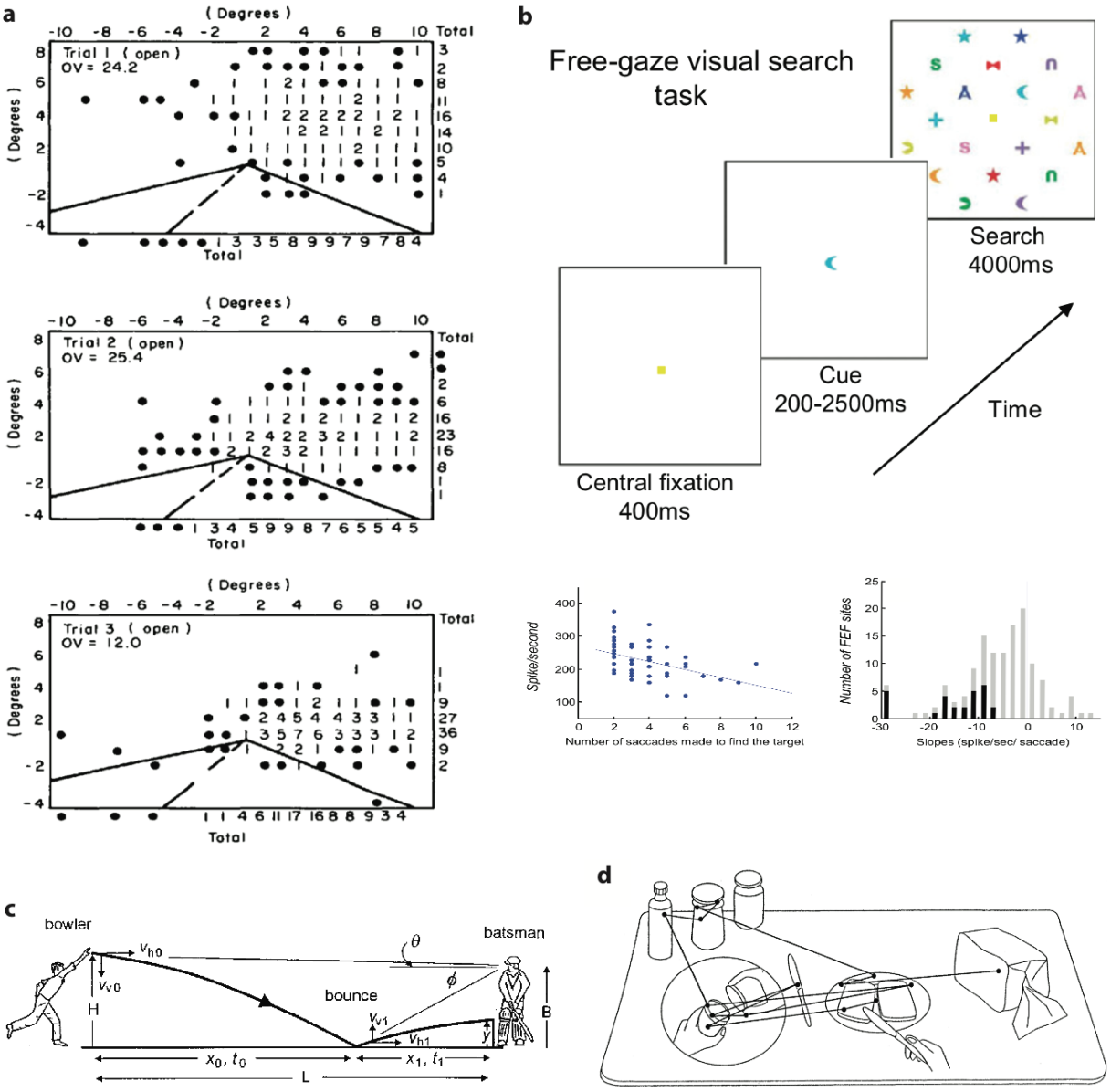


Figure 2.6: Experimental motivation to explore spatial biases, feature biases, and more complex semantic top-down models. **(a)** Percentage of fixations onto different road locations while human drivers drove at 50 miles/hour on an open road three times (once each panel; dots indicate non-zero percentages smaller than 1). We can see that over the three repetitions (top to bottom panels), eye fixations started clustering more tightly around the left side of the horizon line (from Mourant & Rockwell, 1970). This motivates top-down models to learn over time how a task may induce some spatial biases in the deployment of attention. **(b)** Neural recordings in the frontal eye fields (FEF) as monkeys searched from a cued target among various distractors reveals that the number of saccades the animal made to find the target is negatively correlated with neural firing at the target location around (±50ms) the onset of the first saccade (scatter plot shows examples, one dot per trial, from one recording site, and the negatively-sloped linear regression; histogram shows distributions of slopes over all recording sites, with the significant ones in black and all others in gray). This suggests that top-down models can also exploit biasing for specific features of a search target to attempt to guide attention faster towards the target (from Zhou & Desimone, 2011). **(c)** Eye movement recordings of cricket batsmen revealed that their "eye movements monitor the moment when the ball is released, make a predictive saccade to the place where they expect it to hit the ground, wait for it to bounce, and follow its trajectory for 100-200 ms after the bounce" (Land & McLeod, 2000). This suggests that some knowledge of physics, gravity, bouncing, etc. may be necessary to fully understand human gaze behavior in this more complex scenario. **(d)** Eye movement recordings while making a sandwich are clearly aimed towards the next required item during the unfolding of the successive steps required by the task, with very little searching or exploration (Land & Hayhoe, 2001). Thus, recognition and memorization of objects in the scene is also likely to be required of top-down models to tackle such more complex scenario.

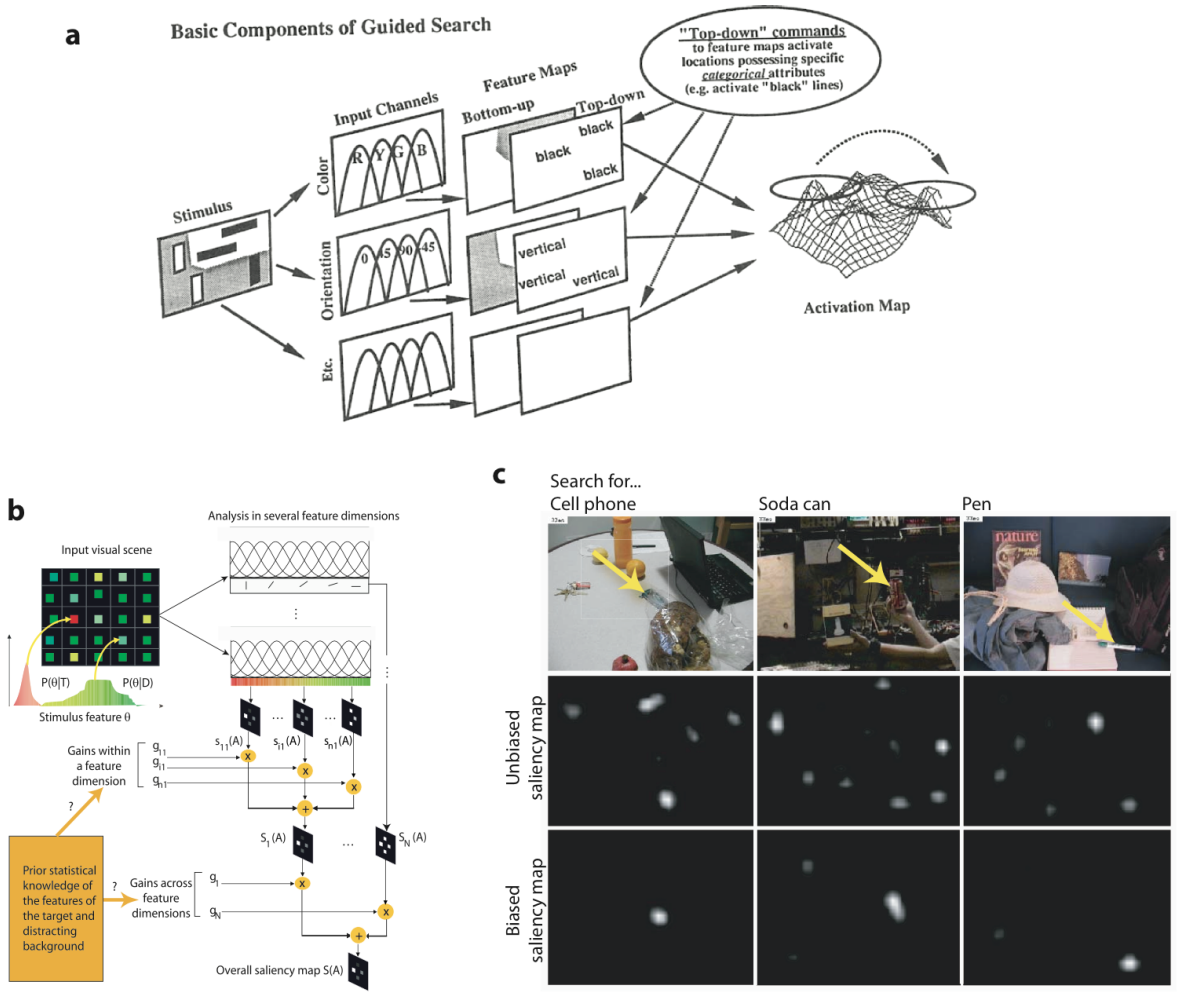


Figure 2.7: Top-down models that modulate feature gains. **(a)** The Guided Search theory of Wolfe (1994) predicted that bottom-up feature maps can be weighted and modulated by top-down commands. **(b)** Computational framework to optimally compute the weight of each feature map, given top-down knowledge of the expected distribution of feature values θ for both target objects ($P(\theta|T)$) and distractors ($P(\theta|D)$). The gain of each feature is set inversely proportionally to the expected target-to-distractor signal-to-noise ratio given these distributions (Navalpakkam & Itti, 2006; 2007). **(c)** Examples of images (top row), naive un-weighted saliency maps (middle row), and optimally-biased saliency maps based on feature distributions gathered from sample training images (bottom row). Although the object of interest was not the most salient according to a purely bottom-up model in these examples, it becomes the most salient once the model is biased using top-down gains computed from the target and distractor feature distributions (Navalpakkam & Itti, 2006).

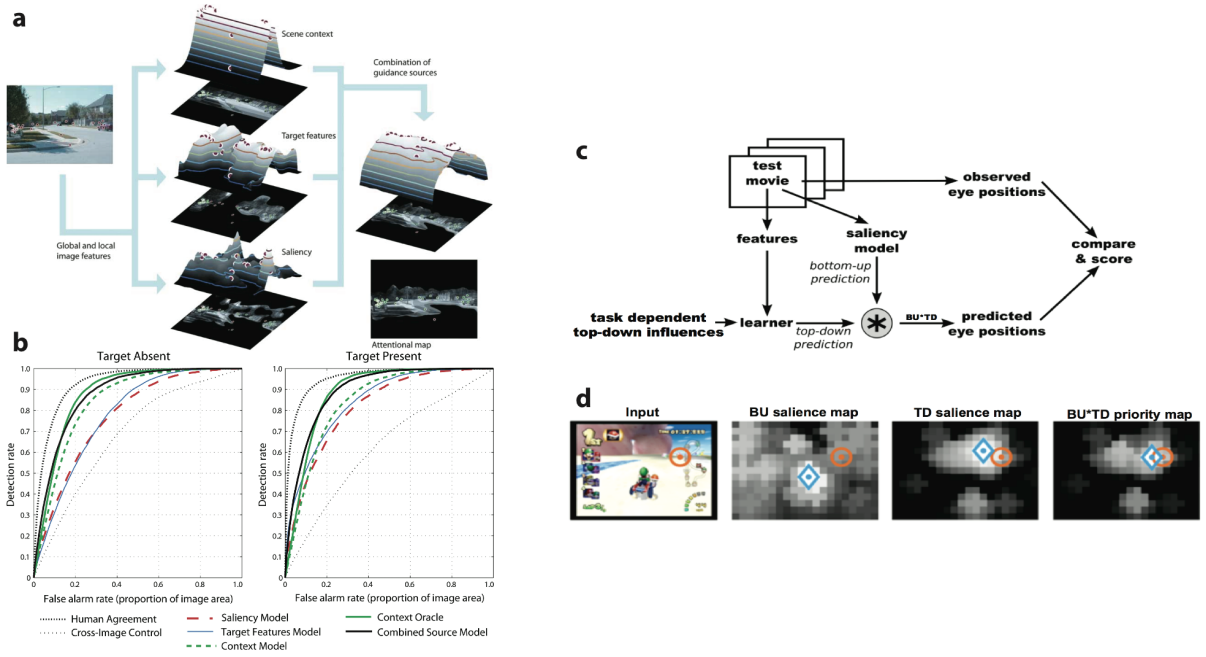


Figure 2.8: Top-down models that involve spatial modulation. **(a)** Model of Ehinger *at al.* (2009) where an attention map is informed by three guidance sources, given the task of finding people: (1) a spatial map that, based on the coarse scene structure, provides a spatial prior on where humans might appear in the given scene (e.g., they might appear on sidewalks); (2) A map that indicates where visual features in the image that resemble the features of the desired targets are observed; (3) a bottom-up saliency map. **(b)** The combined-source model performs best and significantly better than any of the three component models taken alone, and also performs better than an empirical context oracle (where a set of humans manually indicated where humans might appear in the given scenes). **(c)** A model that learns top-down priors from human gaze behavior while engaged in complex naturalistic tasks, such as driving (Peters & Itti, 2007). A task-dependent learner component builds, during a training phase, associations between distinct coarse types of scenes and observed eye movements (e.g., drivers tend to look to the left when the road turns left). During testing, exposure to similar scenes gives rise to a top-down salience map, which is combined with a standard bottom-up salience map to give rise to the final (BU*TD) priority map that guides attention. **(d)** Example results from the model of (c) applied to a driving video game. Blue diamonds represent the peak location in each map and orange circles represent the current eye position of the human driver. Here the bottom-up (BU) salience map considers that the main character is the most interesting scene element, but, as more correctly predicted by the top-down (TD) map, the driver is looking into the road's turn and on the horizon line. From Peters & Itti (2007).

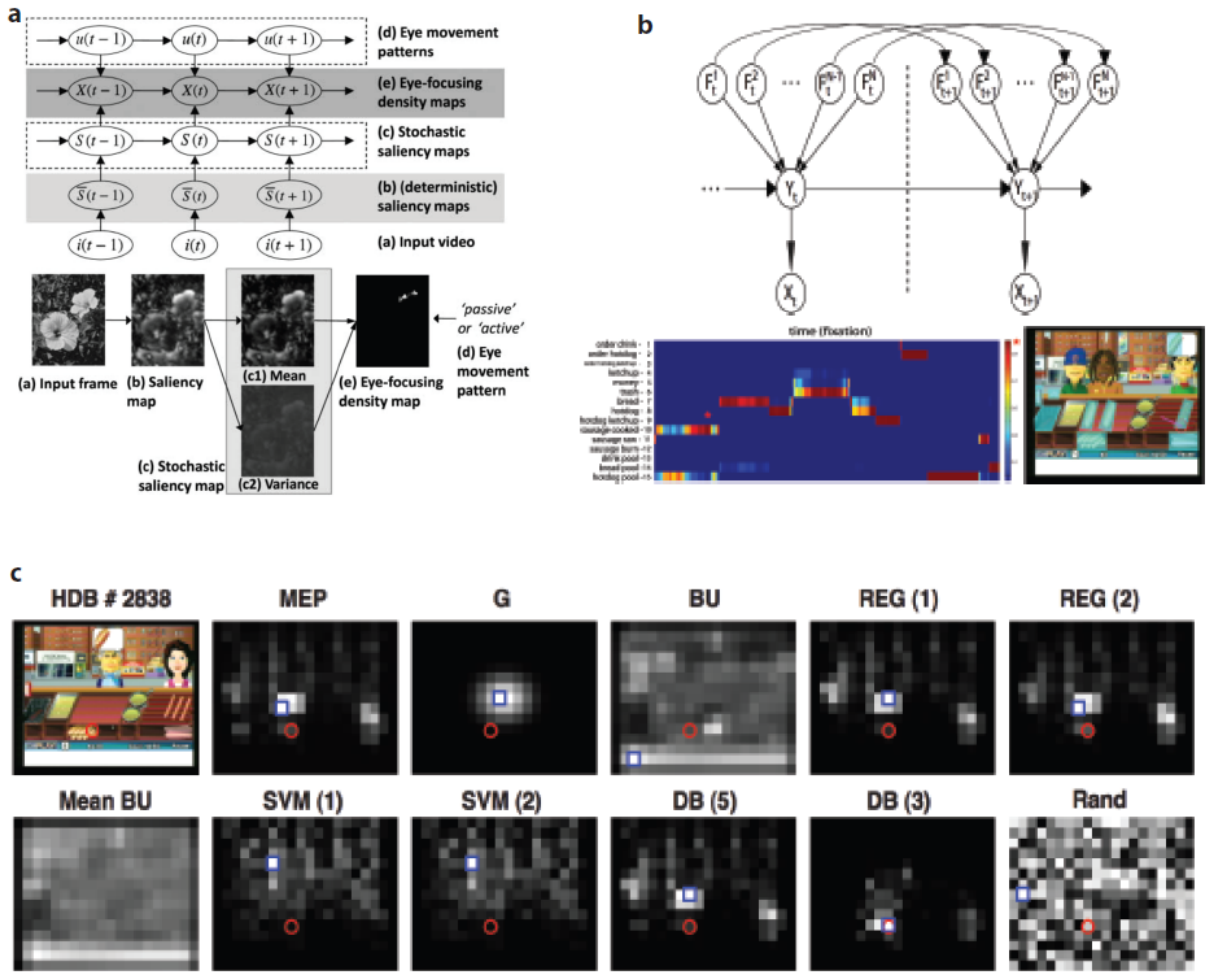


Figure 2.9: Examples of recent more complex top-down models. **(a)** Model of Akamine *et al.* (2012) which combines bottom-up saliency influences and top-down active/passive state influences over space and time using a dynamic Bayesian network. Although the top-down state is quite simple in this model (active vs. passive), the proposed mathematical framework could easily extend to more complex top-down influences. **(b)** Graphical representation of the DBNs approach of Borji *et al.* (2012 AAAI) unrolled over two time-slices. X_t is the current saccade position, Y_t is the currently attended object, and F_t^i is the function that describe object i at the current scene. All variables are discrete. It also shows a time series plot of probability of objects being attended and a sample frame with tagged objects and eye fixation overlaid. **(c)** Sample predicted saccade maps of the DBN model (shown in b). Each red circle indicates the observers eye position superimposed with each maps peak location (blue squares). Smaller distance indicates better prediction. Images from top-left to bottom-right are: a sample frame from the hot-dog bush game where the player has to serve customers food and drink, MEP stands for the mean eye position over all frames during the game play, G is just a trivial Gaussian map at the image center, BU is the bottom-up saliency map of the Itti model, REG(1) is a regression model which maps the previous attended object to the current attended object and fixation location, REG(2) is similar to REG(1) but the input vector consists of the available objects at the scene augmented with the previously attended object, SVM(1) and SVM(1) correspond to REG(1) and REG(2) but using an SVM classifier, Mean BU is the average BU map showing which regions are salient throughout the game course, Similarly DBN(1) and DBN(2) correspond to REG(1) and REG(2) meaning that in DBN(1) network slice consists of just one node for previously attended object while in DBN(2) each network slice consists of the previously attended object as well information of the previous objects in the scene, and finally Rand is a white noise random map.

Chapter 3

Subtle gaze direction, with applications in art and medical image understanding.

3.1 Subtle Gaze Direction

This section has been adapted from Bailey et al. (Bailey, McNamara, Sudarsanam, & Grimm, 2007), (Bailey, McNamara, Sudarsanam, & Grimm, 2009b)

3.1.1 Introduction

This section presents a novel technique that combines eye-tracking with subtle image-space modulation to direct a viewer's gaze about a digital image. We call this paradigm *subtle gaze direction*. Subtle gaze direction exploits the fact that our peripheral vision has very poor acuity compared to our foveal vision. By presenting brief, subtle modulations to the peripheral regions of the field of view, the technique presented here draws the viewer's foveal vision to the modulated region. Additionally, by monitoring saccadic velocity and exploiting the visual phenomenon of *saccadic masking*, modulation is automatically terminated before the viewer's foveal vision enters the modulated region. Hence, the viewer is never actually allowed to scrutinize the stimuli that attracted her gaze. This new subtle gaze directing technique has potential application in many areas including large scale display systems, perceptually adaptive rendering, and complex visual search tasks.

When viewing traditional, static images, the pattern a viewer's gaze makes may be guided by a variety of influences. For example, the pattern of eye movements may depend on the viewer's intent or task (A. L. Yarbus, 1967) (J. M. Henderson & Hollingworth, 1998). Image content also plays a role. For example, it is natural for humans to be drawn immediately to faces or other informative regions of an image (Mackworth & Morandi, 1967). Additionally, research has shown that our gaze is drawn to regions of high local contrast or high edge density (Mannan, Ruddock, & Wooding, 1996) (Parkhurst & Niebur, 2003). Although traditional images are limited to these passive modes of influencing gaze patterns, digital media offers the opportunity for active control of the gaze pattern.

This section demonstrates successful use of subtle image modulation to influence gaze direction, without effectively altering the experience of viewing the image. The technique works by subtly modifying specific locations of the image to direct the viewer's gaze to those locations. This new technique, which combines eye-tracking with subtle image-space modulation, exploits differences in visual acuity and stimuli response time (detection time) between the peripheral vision and the foveal vision of the Human Visual System (HVS).

In humans, the foveal vision has very high acuity compared to the peripheral vision. The falloff in visual acuity as distance from the fovea increases is directly related to the distribution of the cones in the retina (Osterberg, 1935). Fig. 3.1 shows the distribution of cones as a function of angle (relative to the center of gaze). The density of cones, and hence the visual acuity, is very high in the fovea (0 degrees) and

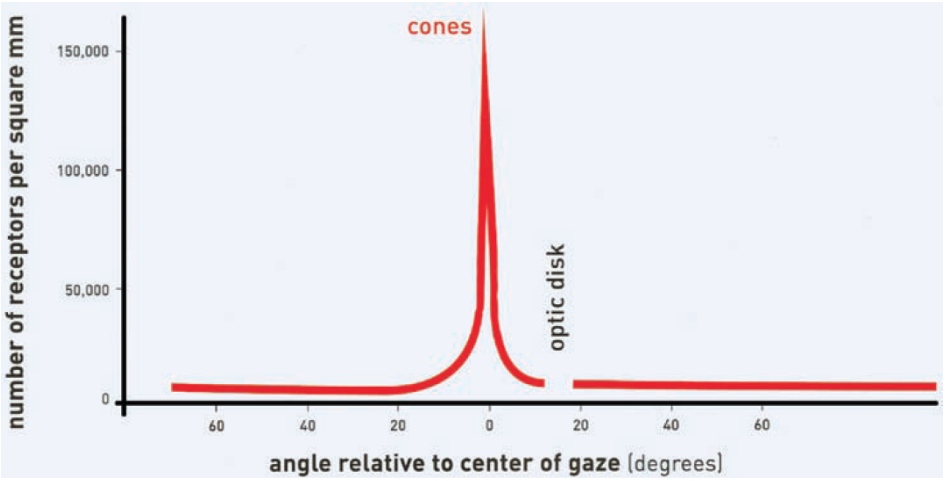


Figure 3.1: Distribution of cones in the retina. Adapted from (Livingstone, 2002). Cones are densely packed in the center of gaze (fovea) and the density of cones falls off rapidly as angle from the center of gaze increases. The distribution of cones directly affects visual acuity. Visual acuity is highest in the center of gaze and falls off rapidly as angle from the center of gaze increases.

falls off rapidly as the angle increases[1]. Research has also shown that the peripheral vision responds (detects stimuli) faster than the foveal vision. This is due to the fact that the optic fibers that carry signals from the peripheral regions of the retina to the primary visual cortex for processing are fast-conducting while the optic fibers that carry signals from the fovea are slower (Ogden & Miller, 1966a).

When viewing a scene for the first time, the low acuity peripheral vision of the HVS locates areas of interest. The slower, high acuity foveal vision is then directed to fixate on these regions. The technique presented in this section operates by modulating regions of the scene that appear only to the peripheral vision. This causes the eyes to move (saccade) to focus the foveal vision on the modulated region in an attempt to identify the stimuli detected. Both luminance modulation and warm-cool modulation were studied. These modulations were chosen because the HVS is very sensitive to luminance changes (Spillmann, 1990), and research has shown that the responses of the photoreceptors in the retina are combined into color-opponent channels that differentiate warm and cool colors (Hurvich & Jameson, 1957).

To achieve subtlety in the gaze directing technique, a small study was conducted to establish thresholds for each type of modulation so that they were just intense enough to be detected by the peripheral vision. This attention capture is the result of the motion cue triggered by the onset of the modulation. However, the viewer's foveal vision is never allowed to fixate on the modulated region. This is achieved by monitoring the direction component of the saccade velocity vector, to determine if the foveal vision is about to enter the modulated region. If this is the case, the modulation is immediately terminated. Taking advantage of the visual phenomenon known as *saccadic masking*[2] allows sufficient time to terminate the modulation before it can be scrutinized by the foveal vision. Saccadic masking, first described by Dodge (Dodge, 1900), is the temporary suppression of visual processing during eye movements (i.e. between fixations (see Fig. 3.2 inset)).

The results of a larger psychophysical experiment reveal that both luminance modulation and warm-cool modulation are effective at directing gaze. Subjective evaluations of the quality of the modulated test images tended to be slightly lower than the corresponding static images. One possible reason for this is that the subtle gaze direction technique forces viewers to violate their natural gaze pattern for a given image.

[1]We ignore the behavior of the rods because our experiment was conducted in a well illuminated (photopic) environment where the response of the rods is completely saturated.

[2]Saccadic masking is also called saccadic suppression. Saccadic masking prevents perception of the blur of the retinal image caused by the ballistic movement of the eye. Complex neurological processes ensure that the resulting gaps in the visual signal are not perceived.

3.1.2 Related Work

Eye tracking systems first emerged in the early 1900s (Dodge & Cline, 1901) (Huey, 1968) (see (Jacob & Karn, 2003) for a review of the history of eye-tracking). Until the 1980s, eye trackers were primarily used to collect eye movement data during psychophysical experiments. This data was typically analyzed after the completion of the experiments. During the 1980s, the benefits of real-time analysis of eye movement data were realized as eye-trackers evolved as a channel for human-computer interaction (Levine, 1981). More recently, real-time eye tracking has been used in interactive graphics applications (D. Luebke, Watson, Cohen, Reddy, & Varshney, 2002) (A. T. Duchowski, 2002) (O'Sullivan, Dingliana, & Howlett, 2003) and large scale display systems (Baudisch, DeCarlo, Duchowski, & Geisler, 2003) to improve computational efficiency and perceived quality. These systems use real-time eye-tracking to *follow* the viewer's gaze. The technique described in this section combines real-time analysis of eye movement data with subtle image space modulation to *direct* the viewer's gaze about a scene.

Up until now, computer graphics approaches for directing a viewer's gaze has typically relied on the fact that the foveal vision is naturally drawn to regions of sharp focus or high detail. The most commonly used approach to direct gaze in 2D images is to simulate the depth-of-field effect from traditional photography. This effect can be achieved in commercially available image editing packages by applying a sharpening filter to specific regions of an image and a blurring filter to others (Mitchell, 2004). This has the effect of bringing different areas of an image in or out of focus. The depth-of-field concept has also been applied to 3D scenes (Kosara, Miksch, & Hauser, 2001). DeCarlo and Santella (DeCarlo & Santella, 2002) used a different approach to direct viewer gaze. They recorded the gaze pattern of a single observer over an image and used it to direct the gaze of others. This was achieved by creating a stylized rendering of the image where only the areas attended to by the first observer are shown in high detail. Cole et al. (Cole et al., 2006) applied similar stylized rendering techniques to direct gaze in 3D scenes. Unlike these previous approaches, which draw the viewer's gaze by manipulating focus or level of detail, the subtle gaze directing technique uses brief luminance or warm-cool image-space modulations presented to the peripheral regions of the field of view. Additionally, unlike the previous approaches, this technique does not affect the overall viewing experience or change the overall appearance of the image.

3.1.3 Subtle Gaze Directing Technique

Consider the hypothetical image shown in Fig. 3.2. Suppose that the goal is to direct the viewer's gaze to some predetermined area of interest A. Let F be the position of the last recorded fixation, let $\vec{v}$ be the velocity of the current saccade, let $\vec{w}$ be the vector from F to A, and let θ be the angle between $\vec{v}$ and $\vec{w}$. Either luminance modulation or warm-cool modulation is performed over the pixels in region A. Once the modulation commences, saccadic velocity is monitored using feedback from a real-time eye-tracking device and the angle θ is continually updated[3] using the geometric interpretation of the dot product:

$$\theta = \arccos\left(\frac{\vec{v}\cdot\vec{w}}{|\vec{v}|\,|\vec{w}|}\right) \tag{3.1}$$

A small value of θ ($\leq 10^\circ$) indicates that the center of gaze is moving toward the modulated region. In such cases, modulation is terminated immediately. It is important to note that the modulation is terminated *during* the saccade to take advantage of the gap in our perception caused by saccadic masking (see Fig. 3.2 inset). This contributes to the overall subtlety of the technique. By repeating this process for other predetermined areas of interest, the viewer's gaze is directed about the scene.

The modulations are simply alternating interpolations of the pixels in A with black and white, in the case of luminance modulation, or with a warm and a cool color, in the case of warm-cool modulation[4] (see Fig. 3.3). Consider a pixel $p = (x, y)$ in A with color $col(p)$. For warm-cool modulation, the color of the resulting pixel $col'(p)$ for the warm interpolation is given by:

$$col'(p) = ((w * i) + col(p) * (1 - i)) * f(p) + col(p) * (1 - f(p)) \tag{3.2}$$

[3]Our software polls the eye-tracker at a rate of 20Hz. This relatively low sampling rate works well in practice because research has shown that only 3-4 fixations occur per second (Gajewski, Pearson, Mack, Bartlett III, & Henderson, 2005).

[4]The alternating interpolations occur at a rate of 10 Hz.

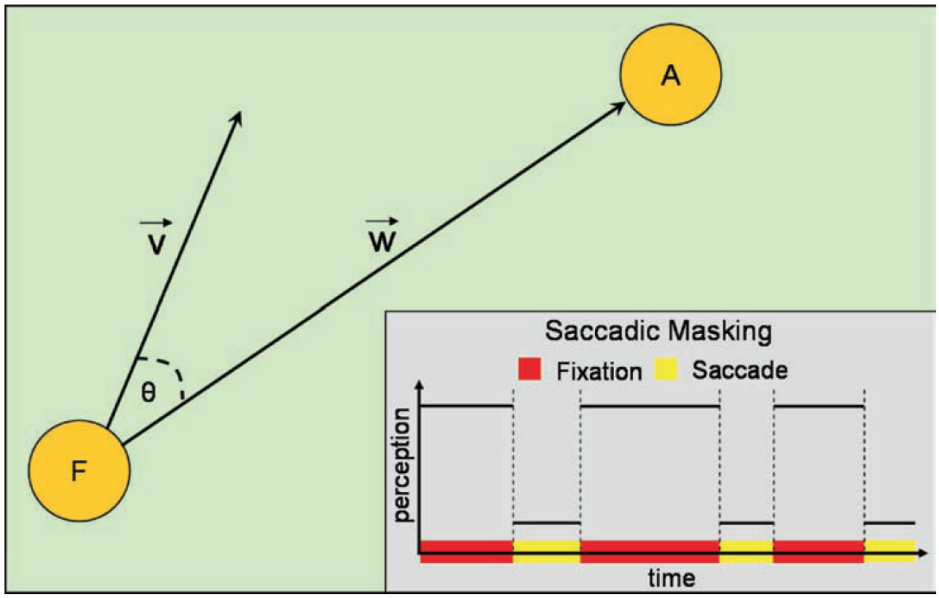


Figure 3.2: Hypothetical image with current fixation region F and predetermined region of interest A. Inset illustrates saccadic masking.

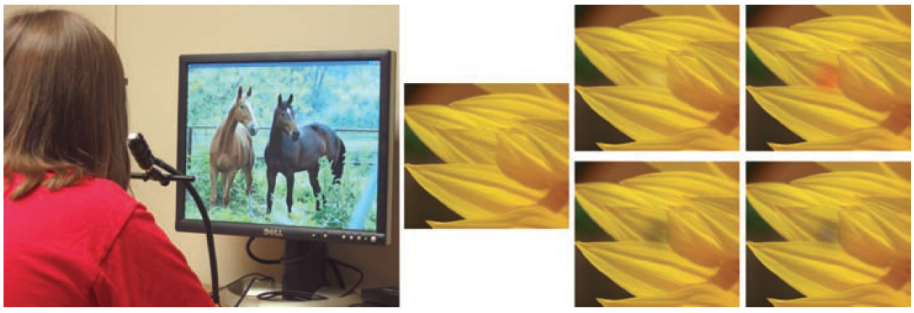


Figure 3.3: Photograph of experiment setup (left). Small patch from a test image (center). Example of luminance modulation (third column). Example of warm-cool modulation (right column).

where w is a warm color[5], $f(p)$ is a falloff function based on the distance of p from the center of region A[6] and i is some scalar value in the range [0,1] which controls the intensity of the modulations. The cool interpolation of the warm-cool modulation cycle, and the black and white interpolations of the luminance modulation cycle, are analogous.

A small pilot study was conducted to determine the value of i for which the modulations are just intense enough to be detected by the peripheral vision. Three participants were involved in this pilot study. They were each presented with five randomly selected images from the complete test set (see Fig. 3.22). They were instructed to fixate on a cross in the center of the image while modulations were presented in random peripheral regions. The modulations were not noticeable to begin with. Using the keyboard (+/-), they adjusted the value i in step sizes of 0.005 until the modulations were just noticeable. The final value for i was obtained by averaging the results of the three participants. For luminance modulation between black and white with a Gaussian falloff, $i = 0.095$ and for warm-cool modulation between red and blue with a Gaussian falloff, $i = 0.105$. It should be noted that i needs to be recomputed if any changes are made to the

[5]For our experiment we use red as the warm color and blue as the cool color.

[6]We use a Gaussian falloff function with a radius of 32 pixels. This corresponds to approximately a $1cm$ diameter circular region on screen.

falloff function or to the colors used for the interpolations.

3.1.4 Experimental Design

This section describes an experiment that was conducted to test the effectiveness of the gaze-directing technique and to study its impact on perceived image quality.

Stimuli

Stimuli were presented on a 20 inch monitor, operating at 75Hz with a resolution of 1280 x 1024. The stimuli used in this experiment consisted of forty 1280 x 1024 images compiled from various sources. These images were chosen with no particular criteria and are shown in Fig. 3.22.

Participants

Ten participants (5 females, 5 males), between the ages of 18 and 45 volunteered to participate in this study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of two groups:

- **Static group:** Participants (five) were presented with a randomized sequence of the 40 images with no modulation. This group served as the control group for the experiment.

- **Modulated group:** Participants (five) were presented with a randomized sequence of the 40 images with modulations at pre-selected image locations. These locations, manually selected by the researchers, included areas that are not visually significant - low contrast, low detail, low color saturation, and uninteresting objects. These are areas that an observer would not ordinarily attend to. The type of modulation for a given image was randomly selected to be either luminance or warm-cool.

  To ensure that the onset of a modulation does not occur in the immediate vicinity of the current fixation or in the path of the current saccade, the eye-tracker is polled just before a new location is modulated. If the new location is within the foveal view the modulation is skipped. To further reduce the likelihood of this occurring, the researchers selected target locations for modulation that were well-spaced across the image.

For both groups, the images were presented for a duration of 8 seconds. Between each image, a black screen with a small white cross displayed at the center was presented for 2 seconds. This allows participants to rest briefly between images.

3.1.5 Procedure

Participants were seated in front of the computer screen in a well lit room with their chin comfortably resting on a chin-rest to reduce head movement. Using an infrared camera-based eye-tracking system[7], dominant eye position was recorded for each participant[8] (see Fig. 3.3). Participants were instructed to remain as still as possible while the eye-tracker was calibrated and the experiment was conducted. The chin-rest was positioned $75cm$ from the screen. At this distance, the actual perceptual span (area of high acuity) of the observer occupies a circular region of diameter $5cm$ on the screen (Rayner, 1975). To further promote subtlety, modulations were presented in a smaller ($1cm$ diameter) circular region.

The participants in each group were asked to assess the quality of each image on a scale from 1 (low) to 10 (high) and to report it verbally during the 2 second period between images. These scores were recorded by the researchers. The term *quality* was not defined by the researchers. Instead, it was left up to the participants to formulate their own notion of image quality. The complete set of instructions read verbatim

[7] ViewPoint EyeTracker® by Arrington Research, Inc.

[8] The dominant eye was established by asking participants to stare at an object a few feet away from them and to point at that object with their index finger. With their eyes focused on the object their index finger would appear blurred in the line of sight. They were asked to close one eye and then the other. The dominant eye is the eye with which the index finger appears to be pointing directly at the object.

Figure 3.4: The complete set of images used in this study. Images were chosen with no particular criteria and gathered from various sources on the web (Yahoo! Inc., 2014) (Google Inc., 2014).

to each participant are given in Appendix **??**. For participants in the modulated group, real-time analysis of the eye-tracking data was performed (as described in section 3.1.3) to determine when to terminate the modulations.

Data files for all participants were compiled during the experiment. The data files contain the following information: time elapsed, filename of image being displayed, position of modulated region, type of modulation, location of current fixation, and corresponding delta change from last fixation. Entries to the data files occurred at a rate of 20Hz. Testing for each participant (including calibration) lasted approximately 15 minutes.

3.1.6 Results

Evaluation of Image Quality

One way to evaluate the effect of gaze-directing modulations on perceived image quality is to compare the mean quality scores for the static and modulated images gathered from the eye-tracking experiment. Fig. 3.5 summarizes these findings. Quality ratings for the static images average 7.07, while the modulated images receive a mean quality rating of 6.17. These mean scores were obtained by averaging the individual scores of the 5 participants over the forty images in each group. This observation suggests that introducing modulations, even subtle ones presented to the peripheral regions of the field of view, results in a reduction in the perceived quality of the image being viewed.

An independent-samples t-test revealed that the effect of the modulation was significant:

$$t(398) = 4.884; p < 0.001$$

This means that the difference in quality judgments between the static and modulated images is due to the presence (or lack) of modulation and cannot be attributed to chance or other factors. We believe that the presence of the modulations causes viewers to violate their natural gaze pattern for a given image. Recall that the modulated regions were specifically chosen because they had no features of interest. Therefore, by drawing the viewer's gaze to these regions, they were left with less time to view parts of the image they would normally attend to. This altered pattern can be clearly seen in Fig. 3.8 and Fig. 3.9, and may be the cause of the lower quality ratings.

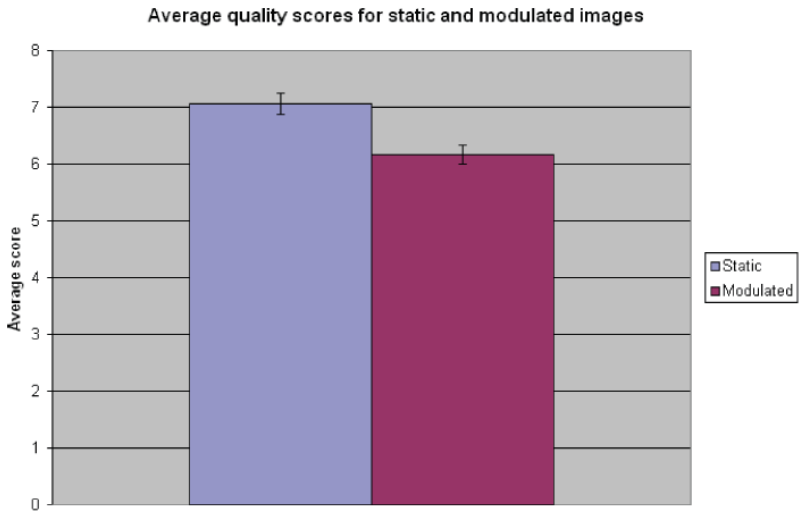
For the purpose of further analysis, we categorize the quality ratings of each static image with respect to the corresponding modulated image. A static image can fall into one of three categories:

- higher perceived quality
- equal perceived quality
- lower perceived quality

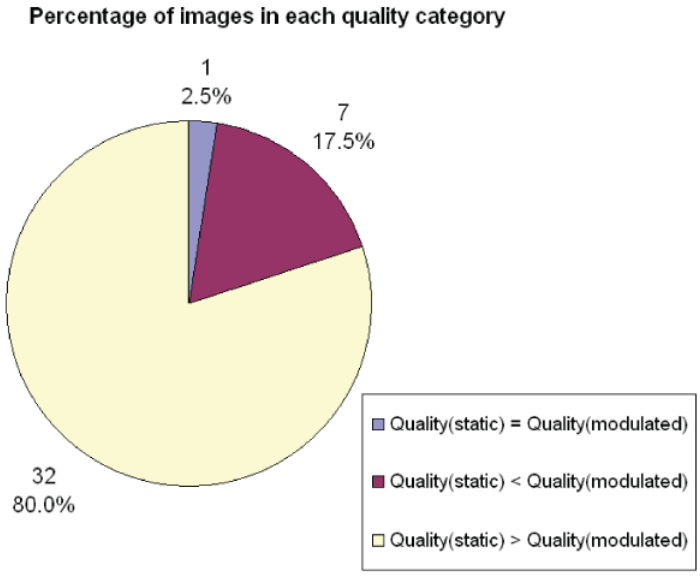
Fig. **??** shows the percentage of images that fall into each of these categories. It is interesting to note that 80% of the static images received higher quality ratings than the corresponding modulated images. On the other hand, only 17.5% of the static images received lower quality scores than the corresponding modulated images. Again, we speculate this is due to deliberate disruption of natural viewing caused by the subtle modulations.

Effectiveness of Gaze Directing Technique

Since only areas of low visual significance were selected as regions of interest by the researchers, it would be expected that an effective gaze directing technique would result in a gaze pattern that is significantly different than one obtain by natural viewing. By comparing fixations in the static scene to those in the luminance modulated and warm-cool modulated scenes, it is possible to evaluate the effectiveness of our gaze directing technique. Fixations denote regions of the image where viewers rest their gaze, indicating a feature of interest in the scene. To facilitate this comparison we determine the total fixation duration in



(a)



(b)

Figure 3.5: (a) Average quality scores for static and modulated images. (b) Percentage of images that fall into each quality category.

various regions of an image in both the static condition and the modulated condition. The regions were specified by partitioning each image into a 5 x 5 regular grid.

For each static image the percentage of fixation time, averaged across the 5 participants, was determined for each grid cell. This averaged fixation time quantifies the natural gaze distribution for a given static image (see Fig. 3.6). Essentially, this data serves as the ground truth for a given image and we would expect any viewer's gaze distribution to correlate highly under similar viewing conditions and assigned task. In fact, to verify this ground truth, a new participant was asked to complete the experiment. As expected, we observed a high degree of correlation between her gaze distribution and that of the averaged distribution. This is especially true for the images with strong salient features where we observed a Pearson coefficient of correlation as high as $r = 0.96$. For images with no dominant salient features, the minimum Pearson coefficient of correlation was still a respectable value of $r = 0.67$. This validates our supposition that averaging over the five participants for the static group would give us a ground truth or normal (average) gaze pattern for free viewing of each image. The challenge now was to demonstrate that gaze patterns over the modulated images did not correlate well with the ground truth indicating that the presence of modulations had caused some upset to the expected natural gaze pattern for a given image.

The participants in the modulated group viewed a total of 200 images (5 participants, 40 images). Of these, 107 were luminance modulated, and 93 were warm-cool modulated (the type of modulation was chosen randomly). Fig. 3.7 plots all the correlation values obtained by comparing the gaze distribution of the modulated images with the corresponding ground truth gaze distribution. The results indicate a poor correlation between the gaze distribution of the two groups. This is to be expected as the technique deliberately attempts to drive the viewer's scan path off their natural viewing path. The poor correlations indicate that viewing the modulated images yields very different results than viewing the static images. This provides an element of confidence that our technique does indeed function correctly and draws the viewer's gaze to pre-selected regions of the image, and prevents them from following their instinctive gaze configuration.

Further statistical analysis is necessary of course to validate that this variation in gaze patterns did not happen by chance. We conducted a repeated measures between subjects analysis of variance (ANOVA) test on the coefficients of correlation using the participants as the factor for the test. The following results were obtained:

$$F(39, 1.903) = 21.034; p < 0.001$$

This shows that the poor correlation values did not happen by chance. Hence we can conclude that the presence of gaze directing modulations results in a gaze distribution that is significantly different than the natural gaze distribution.

Notice in Fig. 3.7 that the coefficients of correlation for the luminance modulated images occupy a narrower range ($-0.34 \leq r \leq 0.66$) than the coefficients of correlation for the warm-cool modulated images ($-0.46 \leq r \leq 0.75$). This suggests that the luminance modulation is more effective at directing gaze than the warm-cool modulation. One possible explanation for this is due to the fact that the luminance modulations are visible against a wider variety of image backgrounds whereas the warm-cool modulations are sometimes difficult to detect, especially against dark or greenish colored backgrounds.

Fig. 3.8 and Fig. 3.9 show examples of the averaged gaze distributions for the participants in the static group and the modulated group for two of the images used in the experiment. These figures demonstrate that our gaze directing technique does indeed result in significantly different gaze distributions, compared to the gaze distributions for the corresponding static images.

Notice that the overlaid gaze distributions are not necessarily well aligned with the researcher pre-selected locations for modulation. There are two possible causes for this inaccuracy. First, the simple chin-rest that we use does not completely eliminate head movement. Second, since the subtle modulations are terminated as soon as the eye starts moving toward them, it may be that the presentation time of the stimuli is to too brief for the brain to accurately determine its location.

On completing the experiment the participants were asked if there was anything unusual about the images. Surprisingly, only one participant reported seeing anything atypical in the image. However, even on noticing the presence of irregularities the participant reported ignoring them and completing the experiment regardless. None of the participants reported anything which would cause confusion or distraction during image viewing. This requires further studies to verify that this is indeed the case and that task performance

(a)

0.14%	2.99%	2.28%	1.59%	0.00%
1.27%	9.60%	6.07%	11.88%	5.98%
0.71%	7.49%	14.98%	4.67%	3.12%
0.21%	1.13%	5.39%	2.19%	2.77%
0.38%	2.84%	4.55%	5.09%	2.68%

(b)

Figure 3.6: (a) Source image. (b) Average fixation time for each grid cell for participants viewing image without modulation. Notice that regions containing visually significant features such as faces receive greater fixation time.

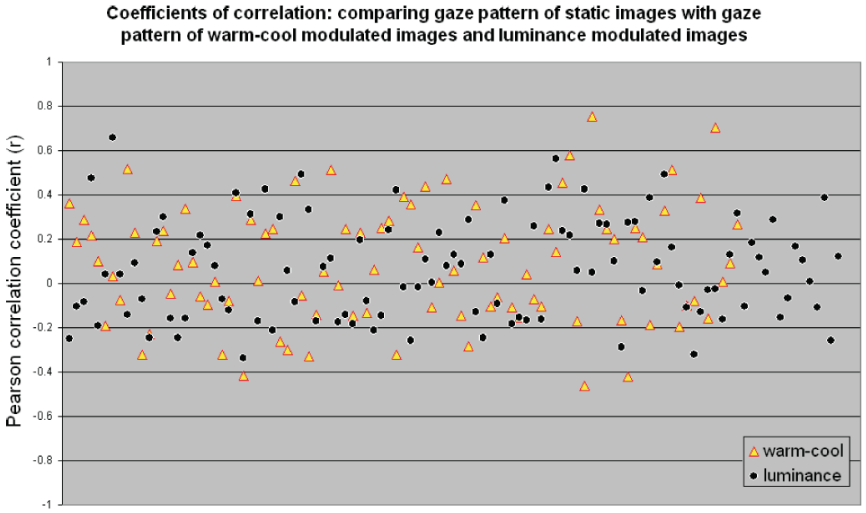


Figure 3.7: Plot of all Pearson coefficient of correlation values obtained by comparing the gaze distribution of modulated images with the corresponding natural gaze distribution. Note that most of the coefficients fall in a very narrow range close to zero ($-0.2 \leq r \leq 0.4$). These small values show that there is poor correlation between the gaze patterns of the static group and modulated group.

(for example) would not be impeded as a result of gaze direction.

3.1.7 Analysis of Activation Times Recorded During Experiment

Recall from our earlier discussion that the viewer's saccade velocity is monitored in real-time and that modulations are terminated whenever the saccade is directed toward the target region. We define *activation time* as the time elapsed between the onset and termination of a modulation. Fig. 3.10 shows the activation times recorded during the experiment for all participants in the modulated group. As expected, the plot of the sorted activation times reveals a stair-step function. This is due to the fact that activation time (as defined) is directly related to the incremental number of saccades that occur between the onset and termination of a modulation. This plot shows that the criteria for terminating the modulation was met within 5 saccades (roughly 0.5 seconds) for approximately 75 percent of the target regions and within 10 saccades (roughly 1 second) for approximately 90 percent of the target regions. This observation provides further indication that the viewers do attend to the modulated regions shortly after their onset.

3.1.8 Conclusion

This section presented a method capable of directing an observer's gaze to chosen regions of a display. Using custom built software, we introduce brief subtle luminance or warm-cool modulations in still images. By monitoring the saccadic velocity of the dominant eye of the observer, we are able to terminate these modulations before the observer has an opportunity to scrutinize them. The results of an experiment show that this approach to gaze directing is highly effective. Subjective evaluations of image quality, however, tended to be lower for the modulated images compared to the corresponding static images.

This technique has potential applications in a number of research and commercial areas:

- **Perceptually Adaptive Rendering:** By dictating where in a scene a viewer looks, we can reap enormous benefits in the rendering domain. If a particular region of a scene requires more rendering time (due to the complexity of the model or the need for higher resolution), our technique can be used to direct users to look at other regions of the scene that required less time to render, thus distracting their view from the more complex areas that are progressively updating. To fully make use of this idea, we would have to gather additional information on how long a user is distracted by a modulation

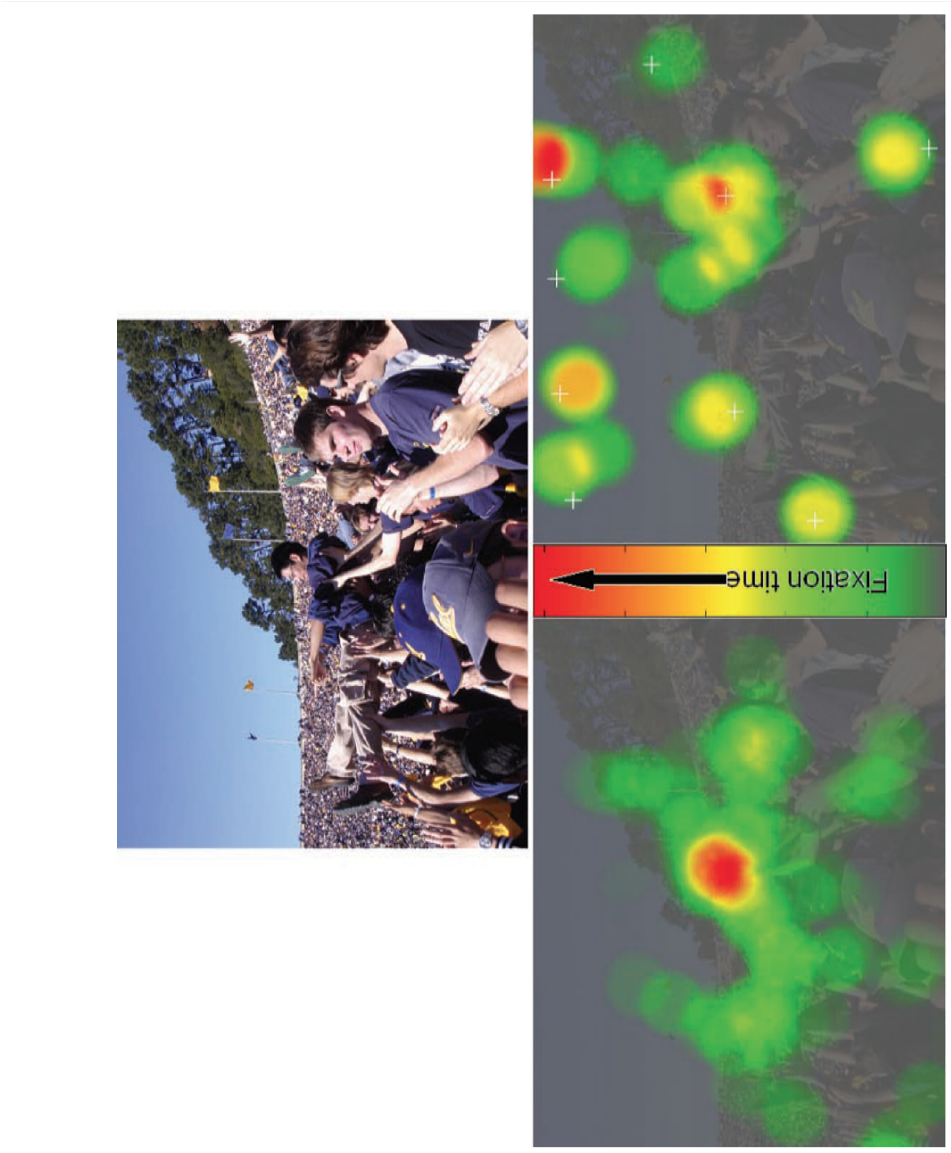


Figure 3.8: Gaze distributions for an image under static and modulated conditions. Input image (top). Gaze distribution for static image (bottom left). Gaze distribution for modulated image (bottom right). White crosses indicate locations preselected by researchers for modulation.

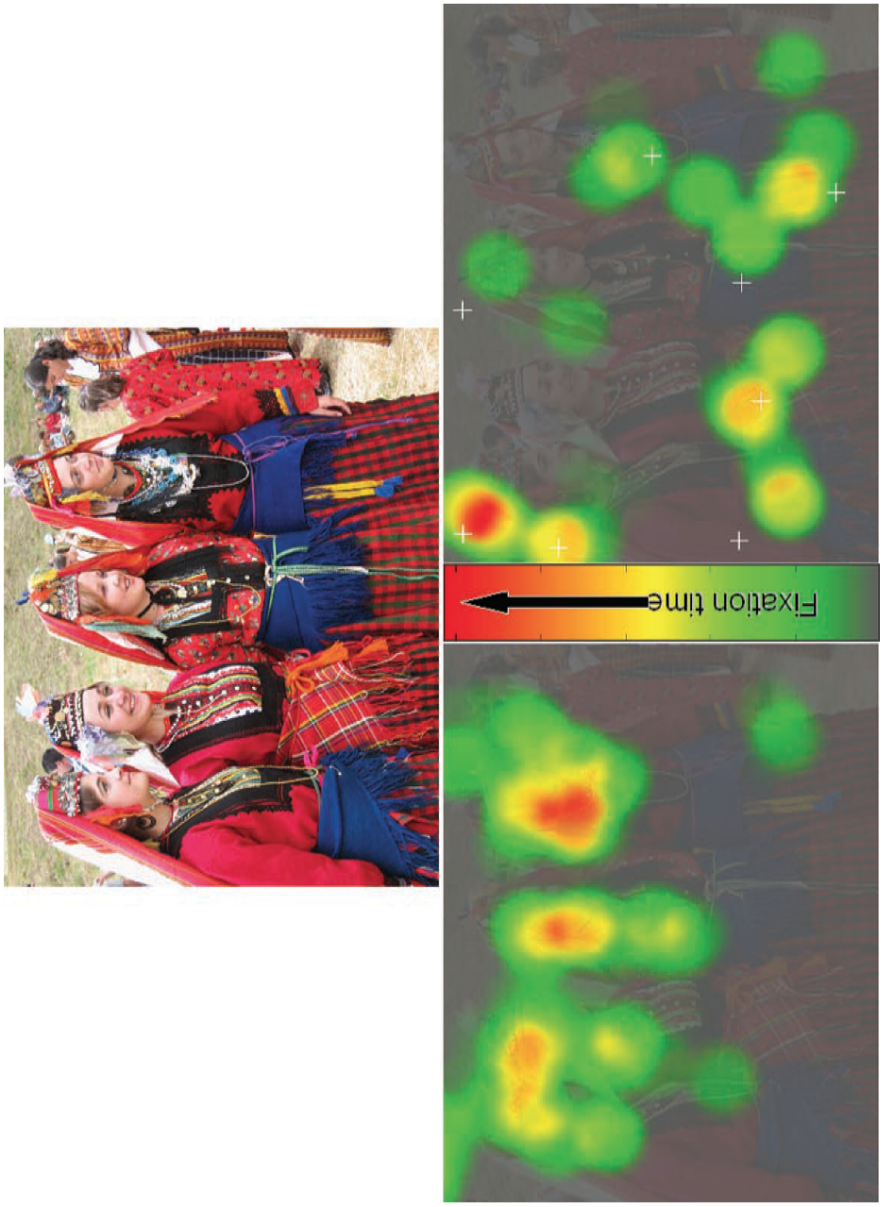


Figure 3.9: Gaze distributions for an image under static and modulated conditions. Input image (top). Gaze distribution for static image (bottom left). Gaze distribution for modulated image (bottom right). White crosses indicate locations preselected by researchers for modulation.

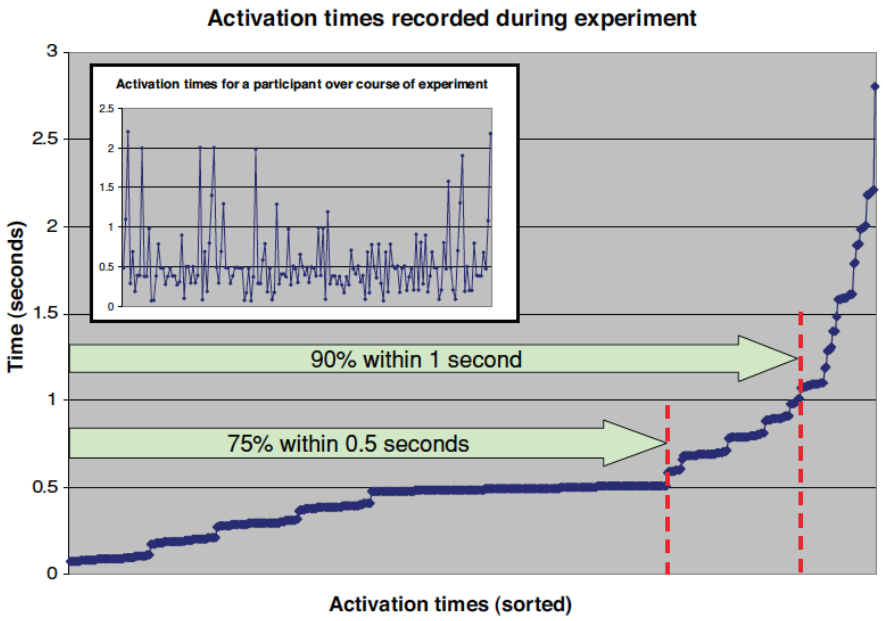


Figure 3.10: Activation times recorded during experiment. The activation times are sorted by increasing time. This graph reveals that approximately 75% of the target regions were found within 5 saccades and approximately 90% were found within 10 saccades.

and how strong the modulation must be in order to distract from changes elsewhere in the image. By coordinating modulations with image updates it might be possible to mask the updates with saccades.

- **Flight/Driving Simulation and Training:** In flight and driving simulators, the goal is to instill good navigation habits, such as checking certain information on the cockpit equipment, or routinely checking rear-view and side mirrors. Training simulators could be equipped with subtle gaze direction techniques to encourage users to frequently look at selected regions, such as mirrors, in the hope that the habit would transfer to the real world situation. Because the imagery in this case is dynamic, additional studies are necessary to determine if, and what kind of, modulations are sufficiently strong to attract the viewer's gaze without decreasing their performance.

- **On-line Training and Distance Education:** Educators employing on-line technology in distance learning courses could use subtle gaze direction to encourage student viewing of relevant sections of the on-line course display. For example, when displaying slides with a voice over, subtle gaze direction could be used to guide viewers to look at relevant text or imagery on the screen. This would involve carefully synchronizing the modulations to reflect the content in the audio.

- **Pervasive Advertising:** On large single-screen displays, advertisers could use subtle gaze directing to quickly guide the viewer to the important product information. This could potentially be a cost saving approach especially for high viewer volume segments such as Super Bowl commercials which currently cost about $2.4M for 30 seconds of air time.

Related to all these applications is the question of image understanding and information recall. Just because a viewer looks at a given image location does not necessarily mean that they fully processed the visual details or remembered them. Further studies would probe just how much information is retained. These studies would need to be content and/or application specific.

There are also several potential avenues of future research. We believe that the effectiveness of the gaze directing technique would be improved by making it more adaptive to the observer's viewing configuration.

For example, by adjusting the intensity of the modulation and/or the size of the modulated area based on the distance between the current fixation point and the desired fixation point, the modulations can be made more salient for larger distances or more subtle for smaller distances. For larger distances, adapting the modulations in this manner will lead to quicker detection by the peripheral vision and also a greater degree of accuracy when the eye saccades to fixate on the modulated region. We also plan to experiment with other types of modulation such as sharpness, contrast, texture, and the introduction of noise.

An emerging area of research that explores the use of subliminal cues to trigger unconscious neural processing may offer a solution to the problem of reduced perceived quality that was observed for the modulated images. The approach that is currently being used presents subtle stimuli to the peripheral regions of one eye while a continuously flashing mask is presented to the other eye. The presence of the flashing mask completely suppresses the conscious perception of subtle stimuli. However, the subtle stimuli still results in electrical signals that trigger neural processing (Bahrami, Lavie, & Rees, 2007). While this particular approach is impractical for our work, because we would like to facilitate natural binocular viewing of the display, it does encourage the exploration of subliminal approaches for gaze-directing.

Finally, a natural extension of the research presented in this section is the application of the gaze-directing technique to video. We believe that the presence of motion in the video will further suppress the subtle modulations. Hence, the modulated videos may not suffer from the perceived quality degradation that was noted in the modulated images. There are however, several challenges such as maintaining frame-to-frame coherency of the modulations and changing the position of the modulations to follow moving regions of interest, that will need to be overcome for this to be feasible.

As display technology advances and screen sizes continue to increase, there is a danger that viewers will be overwhelmed by the amount of visual information available at any one time. If their gaze is subtlety directed as they navigate through the vast amount of information, the effects will be positive for both viewer and presenter.

3.2 Gaze Direction in Search Tasks

This section has been adapted from McNamara et al. (McNamara, Bailey, & Grimm, 2008a)

3.2.1 Introduction

Images provide an effective form of communication. However, visual processing is a complex task that is not yet well understood. A large body of research exists which examines the way in which human observer and process visual information. Eye tracking offers a unique way to study how the Human Visual System (HVS) interprets and organize visual information (Jacob & Karn, 2003) In recent years eye-tracking has been employed to study visual behavior in various research domains including image scanning, driving, solving arithmetic problems, and reading (Hallowell & Lansing, 2001) (Sodhi et al., 2002b). Eye-tracking offers a unique way to study how human observers search and process visual information by monitoring where they look at in a scene. This information in-turn can be used to improve the viewer's visual experience. For example, when rendering complex geometry, fixation data can be used to determine the level of detail necessary for each object (C. H. Lee, Varshney, & Jacobs, 2005) (D. Luebke & Erikson, 1997) (A. T. Duchowski & Çöltekin, 2007) (Murphy & Duchowski, 2007). Although numerous studies have used eye-tracking to follow eye movements, relatively few studies have attempted to use eye-tracking to drive or direct eye movements. Subtle gaze direction (Bailey, McNamara, Sudarsanam, & Grimm, 2008) (Bailey et al., 2007) uses brief, subtle image-space modulations presented to the peripheral regions of the field of view to direct a viewer's gaze about a scene. The technique utilizes eye-tracking to monitor a viewer's saccadic velocity. This information is used to ensure that the viewer is never allowed to fixate on a region that is being modulated.

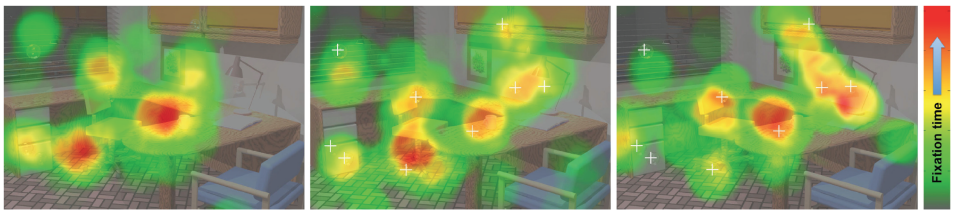


Figure 3.11: From left to right: distribution of fixation time under normal viewing conditions, using Subtle Gaze Direction modulations, and obvious modulations

Humans routinely perform visual search tasks such as searching for a familiar face in a crowd or scanning a document for some important information. Although such searches are a natural part of our visual processing, there are situations in which the task becomes quite complex and demanding. There are numerous factors which impact the difficulty of a visual search. For example, size of the scene, size of the target, subtlety of the target, contrast, number of objects in the scene, etc. Some types of visual searches may even require specialized training and significant experience in order for the viewer to become proficient. In the medical profession, for example, deciphering x-rays while searching for abnormalities is a demanding search task (Schwaninger, Michel, & Bolfing, 2007) (Schwaninger, Hardmeier, & Hofer, 2004).

One way to improve performance in such tasks is to develop a technique to guide the viewer's gaze toward the regions of a scene that are important for successful completion of the task. To date several researchers have focused on *following* the viewer's gaze pattern to gain efficiencies in rendering and presentation (D. Luebke et al., 2002), (A. T. Duchowski, 2002) (O'Sullivan et al., 2003). However, research is beginning to emerge which looks at *directing* a viewer's gaze about a scene (Kim & Varshney, 2006), (Mitchell, 2004), (Kosara et al., 2001), (DeCarlo & Santella, 2002). This paper focuses on the simple task of counting targets in an image (see Figure 3.12). Accurately counting targets efficiently is a necessary task for many applications. For example air traffic controllers need to accurately monitor all aircraft in their vicinity. The gaze directing technique used in this paper, which we call Subtle Gaze Direction, combines real-time analysis of eye movement data with subtle image space modulation to *direct* the viewer's gaze towards selected targets of known location.

One might argue that if the information regarding important regions is available why not simply present that to the user. We agree that in many cases this would be the correct solution, however, cases exist where

Figure 3.12: Example of an image used in this study. The search targets are the transparent spheres "bubbles" in the image.

an algorithm can be beneficial in "suggesting" places to look without disturbing the visual experience of the viewer.

Subtle Gaze Direction depends on the well established fact that the peripheral vision processes stimuli more quickly than the foveal vision (Ogden & Miller, 1966a). When viewing a scene for the first time, the low acuity peripheral vision of the Human Visual System (HVS) locates areas of interest. The slower, high-acuity foveal vision is then involuntarily directed to fixate on these regions. By modulating regions of the scene that appear only to the peripheral vision we can force the peripheral vision to locate areas of interest, which are subsequently focused on. This causes the eyes to move involuntarily (saccade) to focus the foveal vision on the modulated region in an attempt to identify the stimuli detected. Luminance modulations were chosen because the HVS is very sensitive to luminance changes (Spillmann, 1990).

The modulations are made by alternately blending the pixels in a small area with some amount of black, then some amount of white. The modulation blends from black to white at a rate of 10 Hz. We use a Gaussian falloff function with a radius of 32 pixels, which in our setup corresponds to approximately a $1cm$ diameter circular region on screen.

A small pilot study was conducted to determine the amount of black/white for which the modulations are just intense enough to be detected by the peripheral vision. Three participants were involved in this pilot study. They were each presented with five randomly selected images from the complete test set. They were instructed to fixate on a cross in the center of the image while modulations were presented in random peripheral regions. Using the keyboard (+/-), they adjusted the black and values in step sizes of 0.005 until the modulations were just noticeable. The final values (0.095 percent maximum of black/white) were obtained by averaging the results of the three participants.

Additionally, the viewer's foveal vision is never allowed to fixate on the modulated region. This is achieved by monitoring the direction component of the saccade velocity vector, to determine if the foveal vision is about to enter the modulated region. If this is the case, the modulation is immediately terminated.

Figure 3.11 illustrates the results of introducing modulations into an image on the viewer's gaze pattern. The Figure shows a heat map of the average scan patterns over the image for 6 observers. The image on the left shows the scan pattern resulting from normal viewing. The middle and right images were altered by adding modulations at the target regions indicated by white crosses. In the middle image we applied the subtle modulations, in the right image larger modulations were used.

This paper presents a psychophysical experiment that explores the impact of Subtle Gaze Direction on performance during a visual search task. The results show that this method works well without introducing noticeable artifacts into the image.

Figure 3.13: Scenes used in the experiments. All images were 693 by 1024 pixels except the bathroom scene and the interior scene which were 691 x 1024 and 797 x 1024 respectively.

3.2.2 Experiment

Twenty-four images served as stimuli for the experiment (see Figure 3.13). Six environments were chosen and populated with four different target counts, ranging from 4 targets to 12 targets for a total of 24 images. The targets were small transparent spheres roughly uniformly distributed within the scene. Some spheres were deliberately placed so as to be difficult to resolve. The reason for this was to allow us to investigate if those hard to see targets were more easily resolved using modulation. All of the models used to create the images were taken from the RADIANCE web site (Ward, 2000). Presentation order was randomized to eliminate any learning effects. Images were presented for 14 seconds. A black screen with a white cross at the center was presented between each image to allow the participant to refocus on the center of the screen. This also means the initial viewing position for each image is the same i.e. viewing begins in the center of the images. An example of an image used in this study is shown in Figure 3.12. Image sizes varied as shown in Figure 3.13. In cases where image size was smaller than the viewing screen the image was centered on a black background.

Participants were seated in front of the computer screen in a well lit room with their chin comfortably resting on a chin-rest to reduce head movement. Using an infrared camera-based eye-tracking system[9], data pertaining to fixation position and saccades were recorded for the dominant eye of each participant. A fixation is defined as any pause in gaze $\geq 150ms$. Participants were instructed to remain as still as possible while the eye-tracker was calibrated and the experiment was conducted. The chin-rest was positioned $75cm$ from the screen. At this distance, the actual perceptual span (area of high acuity) of the observer occupies a circular region of diameter $5cm$ on the screen (Rayner, 1975). The subtle modulations were presented in a smaller ($1cm$ diameter) circular region.

Eye-tracking was employed to record the viewer's fixation and saccades while counting targets in the various images. Eye-tracking information also served as input to trigger the modulations on targets that were not attended to, in an effort to highlight them so the user could identify them and include them in their count. Image complexity varied as did the number of targets. The behavior of the targets was also varied as follows:

- **Group 1, No Modulation:** no behavioral actions applied to the targets, so images were viewed normally with no modulations.
- **Group 2, Subtle Modulation:** subtle image modulations were used to highlight the target regions in an effort to aid in counting. Modulation was never applied to targets while they were being directly viewed. Any modulation was applied in the periphery only, and modulations were terminated as the user moved their gaze toward the modulated region. Thus the viewer was never allowed to directly view the modulated region. A modulation radius of 0.04 degrees of visual angle was defined to ensure subtlety.
- **Group 3, Obvious Modulation:** subtle behavior was exaggerated so that the modulations were clearly visible by increasing the size of the modulation. Modulation was similar to the modulation applied in Group 2, however, in this condition the modulations were deliberately set to be more obvious. A modulation radius of 0.125 degrees of visual angle was defined to ensure visibility.

The targets subtended visual angles ranging from 0.05 to 0.08 degrees, depending on their location in the scene. Therefore subtle modulations subtended ≤ 0.5 the size of the targets, while the obvious modulations subtended a visual angle of between 1.5 to 2 times the size of the target.

Eighteen participants were assigned randomly to one of the three groups. Participants volunteered from a group of undergraduates. All had normal or correct-to-normal vision and were naive to the purpose of the experiment. Participants viewed the images on a 22" LCD Screen at a resolution of 1200 X 1600 from a distance of 75cm. Head position was held constant using a chin rest for support. The eye movements of each participant were recorded along with a count of the targets found and the time to respond for each image. An informal exit interview questioned the participants about the quality of the images to determine if the modulations in conditions 2 and 3 were disturbing to the viewer. Participants in condition 2 reported nothing unusual, whereas in condition 3 participants reported seeing the modulations.

[9] ViewPoint EyeTracker® by Arrington Research, Inc.

The task involved viewing each scene and counting the number of targets present. Participants verbally reported the number of targets counted on completing the task.

3.2.3 Results and Discussion

Image	None	Subtle	Obvious
Image A: Soda Hall	25.0%	54.2%	66.0%
Image B: Conference	79.2%	79.2%	66.7%
Image C: Interior	12.5%	29.2%	58.3%
Image D: Office	20.8%	62.5%	54.2%
Image E: Bathroom	37.5%	50.0%	29.2%
Image F: Counter	70.8%	62.5%	66.7%
Average	40.97%	56.25%	56.94%

Table 3.1: This table shows the percentage of accurate detection of all the targets in an image. 100% means all of the targets were found. Each column is the average over 4 cases. Standard Deviations were of the order of 2%.

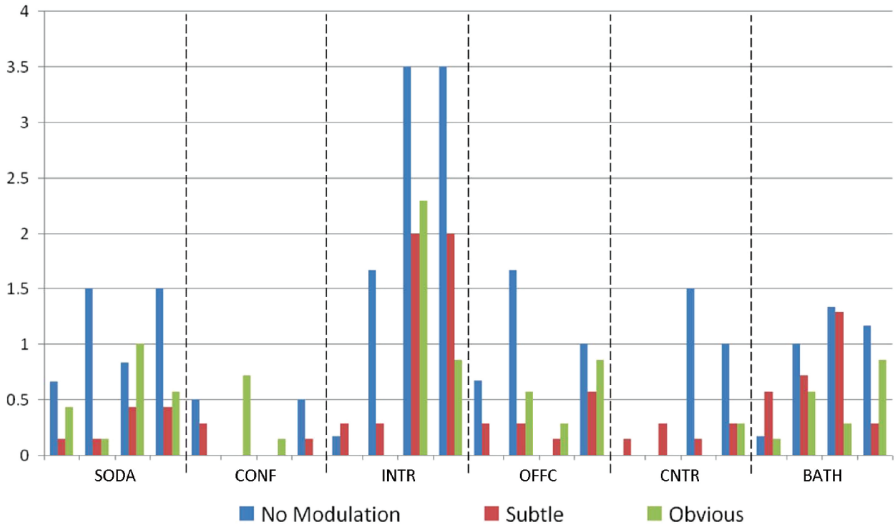


Figure 3.14: Experimental Results: This chart shows the sum of differences between the actual number of targets and the number of targets reported for each condition.

The number of targets reported and the time to respond was recorded for each image. The average response times were consistent across all three conditions, 6.272, 6.495 and 6.570 seconds (with standard deviations of 0.55, 0.94 and 0.96) for the no modulation, subtle modulation and obvious modulation conditions respectively.

We compared the reported number of targets to the actual number of targets and used this to define a correlation. The correlation values represent how close the reported number of targets were to the actual number of targets. Higher correlation corresponds to more accurate task performance. Correlation is higher for the modulated conditions than in the static image with values of 0.80, 0.89 and 0.90 for groups 1, 2 and 3 respectively. This indicates that participants did slightly better on task with the aid of modulation as opposed to normal viewing.

Another interpretation of the results is to simply compare the number of targets missed during counting in each case. This data is shown in Figure 3.14. Each bar represents the absolute difference between the actual

number of targets and the number reported. Smaller bars indicate more accurate counts, no bar indicates 100% precision. The blue bars show the differences in the normal no modulation viewing condition, while the orange and green bars show the modulated images, subtle and obvious respectively. The four bars represent the number of targets for each image (the number of targets ranged from four to 12). Each cluster is one image, as labeled on the x-axis. As the data shows, in most images, the modulation aids in the accuracy of the results. In some cases the number of targets reported does not increase as a function of the number of targets, one reason for this may be the participant's failure to see the targets, and subsequent failure to include them in their count.

The percentage of correct counts reported also reveals that a higher percentage of counts returned in the modulated imagery were accurate compared to the imagery with no modulation. What is interesting in this analysis is that the percentage correct in both modulated cases is higher than in the static imagery. However, it is important to note that while no obvious distractions were noted by participants viewing the subtle modulation case, all participants in the obvious modulation condition reported seeing the modulations. They noted that while the modulations did distract their gaze, it also helped them to identify targets they may not have otherwise counted. This indicates that simple region highlighting, even if noticeable, can contribute to improving task accuracy. The data suggests that subtle gaze direction, where the highlighting is sufficiently faint so as to go unnoticed, successfully guides the viewer's gaze to the target regions, thereby improving task performance. Data is tabulated in Table 1.

The highest discrepancies occurred in the image of the interior scene (Image C). Here the composition of the scene may have influenced the visibility of the bubbles, with only one person getting 100% accuracy. The placement of the targets was also made deliberately difficult. One reason for the poor results in the bathroom scene may be due to the fact that participants reported being confused regarding the inclusion or exclusion of targets reflected in the mirrors.

Further statistical analysis of the data was conducted. A between-subjects ANOVA over the correlations resulted in $F(23; 15) = 64.04; p <= 0.001$. This gives evidence that a significant effect of condition is present between the tasks i.e. performance differed in each group. This can also be seen in Figure 3.11 which compares distribution of fixation time across images.

In summary the results show slight improvement in task performance when modulation is employed to direct gaze to target regions. This seems to hold true whether or not the viewer notices the modulations. Some applications may elect to include obvious modulations whereas in other applications subtlety may play a key role.

3.2.4 Extending Simple Search to include Distractors

Introduction

Many vision algorithms are known to return "false positives". In medical applications pattern classifiers and image processing techniques are routinely used to identify information anomalies, for example in cancer diagnosis (GD, R, Jr, & Floyd CE, 2003), (Tu, Zhou, Bogoni, Barbu, & Comaniciu, 2006a), (Tu, Zhou, Bogoni, Barbu, & Comaniciu, 2006b). Data from these images can be hard to read and interpret by the human eye, and so vision techniques are employed to guide diagnosis, (El-Baz et al., 2006), (Mancas, Gosselin, & Macq, 2004). Generally these techniques come with the caveat that the will performs automatic "suspect" localization, feature extraction, and diagnosis of a particular pattern-class. The overall aim is higher rate of of true-positive fraction detection and low false-positive fraction detection. This means that some false-positive information may still be present in the results. However, the information returned by these algorithms is still maintains a high level of usefulness for certain applications. For example in Figure **??** two mammogram films are depicted. The irregular mass on the breast has been outlined by experienced personnel simply drawing on film in the left image. The image on the right shows the results generated using a computer vision techniques. It correctly identifies the breast, but also returns two false positives (GD et al., 2003), (AO, Jr, LW, & JY, 2003).

Outside the medical realm such algorithms are also used successfully. At the airport vision techniques are employed to segment images of luggage in an effort to detect possible threats.

We were interested to see how well Subtle Gaze Direction would perform if the modulations were driven by the results of such an algorithm. To test this hypothesis we reran experiment two (**the subtle modulation group**) from the previous set of experiments i.e. subtle modulations, with the amendment that this time

we included extra modulations which did not pertain to target location i.e. distractors. In this experiment participants were exposed to modulations not only on the targets but also in random locations away from targets i.e. the "false positives".

For this follow on experiment the set up was essentially identical to the original experiment, with some alterations to allow for inclusion of the distractor modulations. The same 24 image were used as the original experiment. This time the modulations appeared not only on the targets but also in random locations away from the targets. Three additional "distractor" modulations were added to each image. Six new participants were included in the study. All had normal or corrected-to-normal vision. Each image was viewed for 15 seconds. Participants reported a count on the number of targets in the images. Fixations were recorded while the counting task was performed.

Experimental Results

Taking a correlation between the actual number of targets and the number of targets reported gives an initial metric of search performance in the images which were modulated with additional distractors. The average correlation over all images was .93 (and ranged over participant from .89 to .96). These correlations are significantly higher than those reported in the original experiment. Figure **??** shows the correlations for each participant.

As before we report the sum of differences i.e. the difference between the number of targets present and the number of targets reported by the participant. Again the differences i.e. errors were smaller in the distractor condition.

A possible reason for this increase in accuracy is that the distractors attracted gaze and forced viewers to distribute their gaze more wholly over the image. Inattentional blindness has shown that people can miss even very salient features in an image if their attention is focussed elsewhere (Mack & Rock, 1998), (Cater, Chalmers, & Ledda, 2002), (D. Simons & Chabris, 1995). By re-focussing attention, through Subtle Gaze Direction, the viewer is prevented from focussing on a single region for long periods of time and avoid some inattentional blindness by moving their gaze over larger portions of the image.

3.2.5 Conclusions and Future Work

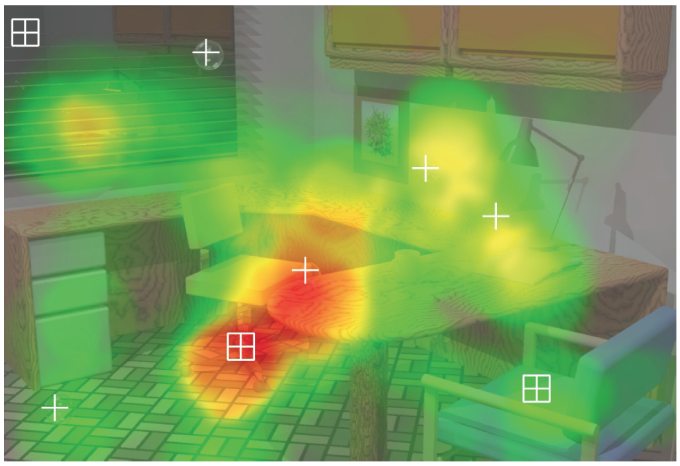


Figure 3.15: Experimental Results: This chart shows fixation distribution for one image in the "distractor" condition. When compared to earlier results it can be seen that image coverage is greater, and this may contribute to greater accuracy on task.

We presented an experiment to compare task performance in digital images across three sets of stimuli. In summary the results indicate that using either subtle or obvious image modulations on the target regions improves the precision of a simple counting task. The difference between using subtle and obvious modulations is the level of disruption to image viewing. With subtle image modulation none of the participants

reported noticing the modulations, whereas with the obvious image modulation all participants reported seeing the modulations.

The success of employing Subtle Gaze Direction on targets to guide viewers in a simple search task led us to investigate the extent to which this success could carry over into images where modulations were presented not only on targets, but also on non-target regions. The reason for investigating the degree to which the technique could improve efficiency in a search task in even with the presence of distractors is important for many reasons. As mentioned earlier many image processing and vision algorithms are known to return "false positives". The presence of which does not render the algorithm useless. Subtle Gaze direction may also find application in interactive environments such as games and animation. Inherently in dynamic scenes such as animation, or navigation through virtual worlds several objects may be moving. To successfully incorporate Subtle Gaze Direction into such environments we need further studies which gauge the performance of the technique in the presence of such "distractors".

In these experiments we focussed only on modulation the luminance channel. It may be that the modulations would be more successful in certain conditions if we used other channels. We have experimented with the warm-cool channel, but found luminance to be slightly more efficient (Bailey et al., 2008). It may be that the most effective modulation should be a function of the image itself, and may be dynamic depending on the behavior in the scene.

The results from this initial study are promising and several follow up experiments are imminent. There are several avenues open for future investigation. In this experiment participants were asked to identify targets and there were no other distracters in the image. Often in visual search the task involves discriminating targets from non-targets (enemy versus friendly for example). In a follow up experiment we showed that the presence of distractors does not distract from task performance, and in fact could be an accurate visual aid due to encouraging people to look over the entire image. It is well known that we as humans suffer from *inattentional blindness* a phenomenon which essentially renders us blind to objects in our visual field which would otherwise be obvious due to the fact our attention is focussed elsewhere. We could use Subtle Gaze Direction to force attention away from the object currently focussed on in an effort to distribute gaze over an entire image and hence provide a more robust image guidance technique.

We have only just begun to explore the potential use of Subtle Gaze Direction in the medical arena. In our next experiments we would like to include real data from medical applications and focus on the detection breast or other cancers, tumors, or other anomalies that may be hard to detect in medical imaging. Coupling Subtle Gaze Direction with a robust vision algorithm, that occasionally returns false positives, could prove an invaluable tool for training and detection of difficult to read information. The results from the vision algorithm could be used to select the modulated regions in the image, and even with false positives (or distractors) present performance on search tasks could potentially be enhanced.

One further possible line of inquiry would be to examine the usefulness of Subtle Gaze Direction in imagery where the task involves identifying or separating targets from non-targets. By applying Subtle Gaze Direction in target regions gaze could be directed only to the targets, making them more distinguishable from non-target regions. Another interesting problem is that of moving targets, or moving imagery generally. Future experimentation will focus on the performance of subtle gaze direction in dynamic environments, such as animations and interactive environments. Subtle Gaze Direction could be used to help guide a user's navigation, or to highlight those parts of an animation that are more relevant to the application. We would expect that stronger modulations would be necessary in order for Subtle Gaze Direction to be effective in dynamic scenes.

We have shown that Subtle Gaze Direction can improve people's performance on a counting task without noticeably changing the image. Example uses of this technique might be to help guide gaze in more complex visual search tasks where targets are numerous or difficult to identify.

Figure 3.16: The cameras and screens in smartphones, tablets and other mobile devices now serve as uniquely convenient tools to combine real world data with virtual data

3.3 Gaze Direction in Art

This section has been adapted from McNamara et al. (McNamara et al., 2012), (McNamara, 2011).

3.3.1 Augmented Reality

During the early 1990s engineers working at Boeing coined the term *Augmented Reality* (AR) to describe systems they developed to overlay computer generated material on top of the real world (P. & W., 1992). Five years later, researchers at the University of California Santa Barbara introduced the "Touring Machine", (Feiner, MacIntyre, Hollerer, & Webster, 1997), the first Mobile Augmented Reality (MAR) system. For the next two decades similar innovations continued to emerge, but it wasn't until many years later, in 2009, that the first AR applications were deployed to smartphones, literally putting AR applicants into the hands of the mainstream.

Advancements in mobile device and AR technology are opening up exciting opportunities. "The cameras and screens in smartphones, tablets and other mobile devices now serve as uniquely convenient tools to combine real world data with virtual data. Sensor-based AR uses GPS capability, image recognition, and the devices' built- in compasses to pinpoint where a mobile device is on the planet and where its camera is pointing, and then uses that information to overlay relevant facts, data, or visuals at appropriate points on the screen" (NMC, 2012). While using smartphones as an AR platform has seemingly limitless potential, there are several hardware issues which require clever research solutions in order to become useful. For example, the limited screen real estate amplifies some of the research challenges posed when implementing AR applications on such a small display. These chapters focus on interface paradigms by developing fundamental algorithms for information presentation in MAR applications. A systematic examination of the placement and saliency of graphic elements which accounts for visual attention will form them basis for new paradigms by aiding scene navigation through novel attention direction techniques.

This section describes the use of Subtle Gaze Direction (SGD) to correctly guide observers through pictorial episodes when viewing paintings. To improve visual literacy we propose the use of SGD to direct the viewer's gaze over an image in a manner which reveals the story in the narrative art. Using a simple ordering task we compared performance using no modulation and using subtle modulation, highlighting the center of the panels in the order of the story narrative. Results from experiments show improved performance when SGD is employed. This experiment establishes the potential of the method as an aid to visual navigation in images where the viewing order is unclear.

3.3.2 Narrative Art

Narrative art tells a story, either as a moment in an ongoing story or as a sequence of events unfolding over time. A synoptic narrative depicts a single scene in which a character, or characters, are portrayed multiple times within a frame to convey that multiple actions are taking place. This can cause the sequence of events to be unclear within the narrative. Synoptic narratives typically provide visual cues that convey the sequence, but it still might be difficult to decipher for those unfamiliar with the story. The process is best illustrated with an example. Figure 3.17 (top) shows a synoptic painting titled *"The Tribute Money"* by Renaissance artist Masaccio. This painting describes a story from the Gospel of Matthew, in which Jesus directs Peter to go to the river and retrieve a coin from the mouth of a fish in order to pay the temple tax. The optimal way to visually navigate this piece is to begin in the center with the tax collector demanding the money. Jesus, surrounded by his disciples, instructs Peter to retrieve the money from the mouth of a fish. By moving their gaze to the *left* of the painting (perhaps counter-intuitive to western civilization who normally read left to right) viewers notice Peter executing Jesus' instruction. The viewer's eyes next need to travel to the extreme right of the painting to view the third episode in which Peter pays the tax collector. At the time it was painted, audiences were conditioned to recognize repeated elements in a frame and identify panels, thereby intuitively understanding the intended order in which each episode of the painting was to be viewed. However, our ability, as artists and audiences, to correctly "read" these paintings may not be so accurate in present day because our visual literacy is not conditioned to follow the viewing pattern the artist intended. In the 15^{th} century the audience would understand that there are multiple episodes in this painting, and also in which order to view these panels in order to comprehend the story. Web-based solutions do exist which manipulate a digital representation of a painting using strong outlines, or interruptive text over the image to explain where the viewer should direct their gaze (see Figure 3.17). While these represent a promising initial approach, a more elegant solution would not disrupt interrupt the visual experience of the audience. Employing gaze direction techniques would allow the viewer to see the actual painting with areas of interest accentuated in a manner which preserves the visual experience by acknowledging the artist's intent. In this work we investigate the use of Subtle Gaze Direction (SGD) as an aid to navigate narrative art. This goal of this work is to satisfy the need to display information in a manner that minimizes disruption to the viewer, but can accurately direct gaze to certain locations of an image, *in a specific sequence.* In other words, our original SGD technique did not examine how well SGD directs gaze to multiple image locations in a specific sequence.

Figure 3.17: Above: *"The Tribute Money"*, by Masaccio tells the story of Jesus and the tax collector. The piece should be viewed in the following order: center, left, then right. Current web-browser based educational tools use text pop-ups with interruptive rectangular outlines to highlight important information in a visual narrative. This not only distracts the viewer from appreciating the image, but also breaks up the image into smaller pieces so it is not viewed in a holistic manner. The red colored rectangle destroys the visual experience by superimposing a distracting overlay on the original painting.

This technique will be especially useful in the digital humanities, as it will allow scholars to recuperate various types of visual literacy specific to a historical moment. We focus on Art History Education as an application but it is easy to extend to any visual task in which viewing order is critical to understanding or task completion.

Imagine a scenario in which an Art History major is trying to improve his visual literacy skills. Narrative

art tells a story, either as a moment in an ongoing story or as a sequence of events unfolding over time. A synoptic narrative depicts a single scene in which a character(s) are portrayed multiple times within a frame to convey that multiple actions are taking place. This can cause the sequence of events to be unclear. Synoptic narratives typically provide visual cues that convey the sequence, but still might be difficult to decipher for those unfamiliar with the story. For example, the student is studying The Tribute Money by Renaissance artist Masaccio, Figure 1. This painting describes a scene from the Gospel of Matthew, in which Jesus directs Peter to find a coin in the mouth of a fish in order to pay the temple tax. The optimal way to visually navigate this piece is to begin in the center with the tax collector demanding the money, Jesus surrounded by his disciples instructs Peter to retrieve the money from the mouth of a fish. By moving their gaze to the left of the painting (perhaps counter-intuitive to western civilization who normally read left to right) viewers notice Peter executing Jesus instruction. The viewers eyes finally need to travel to the extreme right of the painting to view the third episode in which Peter pays the tax collector. At the time it was painted, audiences understood the order in which each episode of the painting was to be viewed to convey the correct story. However, our ability, as artists and audiences, to correctly read these paintings may not be so accurate in present day because our visual literacy is not conditioned to follow the viewing pattern the artist intended. While web-based solutions exist to show the narrative, they manipulate a digital representation of a painting using strong outlines, or interruptive text over the image to explain where the viewer should look, Figure 1. While these represent a good first start, a more elegant solution would not interrupt the visual experience of the audience. Employing mobile AR devices with eye-tracking capabilities would allow the viewer to see the actual painting with areas of interest accentuated in a manner which protects the visual experience. This scenario illustrates the need to display information in a manner that minimizes disruption to the view, but can direct gaze to certain locations of an image, in a specific sequence. Now imagine an AR scenario that accounts for where the user is looking on their mobile device, and delivers content based on gaze location. Not only that, but it delivers that information to an area of the screen that will not obstruct image features that are (or will become) important to the user. Also, imagine a complementary AR system that can influence where viewers look in a scene, both spatially and temporally. This work proposes strategies to realize these AR scenarios. The ideal outcome is an eye-tracking AR system that is fully integrated into mobile devices and can inform AR applications on the optimal placement of AR elements based on gaze information, and also manipulate AR elements to direct visual attention to specific regions of interest in the real world. A healthy number of (mobile) AR applications have successfully been applied in the Art domain. To date, however, few have proposed eye-tracking as an added dimension. The novelty of this approach lies in the eye-tracking and in attracting and directing the gaze to the correct region of the artwork in a sequence that will encourage appropriate visual navigation and understanding of the image and strengthen observation skills

3.3.3 Previous Work

Subtle Gaze Direction (SGD) (Bailey et al., 2009b) exploits the well established fact that human peripheral vision processes stimuli faster than foveal vision (Ogden & Miller, 1966b). On initial viewing of a scene the low-acuity peripheral vision of the Human Visual System (HVS) locates regions of interest. The foveal vision, which is slower and has higher acuity, is then involuntarily directed to focus on these regions. SGD proceeds by modulating regions of an image that appear only to peripheral vision. In this manner peripheral vision is forced to locate the indicated regions of interest, which are subsequently fixated on. This causes involuntary saccades to move the eye to fixate on the modulated region as it attempts to resolve the detected stimuli. Luminance modulation works well as the HVS is highly sensitive to changes in luminance values (Spillman, 1990).

Modulations are constructed by alternately blending small pixel regions with some amount of black, then some amount of white. The rate at which the blend is modulated is 10Hz. A Gaussian falloff with a radius of 32 pixels is used which (in our viewing configuration) corresponds to approximately a 2cm diameter circular screen area.

The marriage of technology and art appreciation is not new — several existing applications have successfully been applied in the Art domain (Gwilt, 2009) (Damala, Cubaud, Bationo, Houlier, & Marchal, 2008) (Andolina et al., 2009) (Bruns, Brombach, Zeidler, & Bimber, 2007) (Choudary, Charvillat, Grigoras,

& Gurdjos, 2009) (Chou, Hsieh, Gandon, & Sadeh, 2005) (Srinivasan, Boast, Furner, & Becvar, 2009). To date, however, few have proposed eye-tracking as an added dimension. The novelty of this approach lies in the eye-tracking and in attracting and directing the gaze to the correct region of the artwork in a sequence that will encourage appropriate visual navigation and understanding of the image and strengthen observation skills.

Obviously conspicuous objects in a scene (such as a black sheep in a white flock) will draw the viewer's attention first. However, there are more subtle image characteristics that can also draw our gaze. Image properties such as color, size and orientation can be used to control attention (Veas, Mendez, Feiner, & Schmalstieg, 2011) (Underwood & Foulsham, 2006). In movies, directors use an arsenal of cinematographic "tricks" to lead the audience to look where they want them to look (see (Bordwell, 2011)). Taking an automated approach, Itti and Koch (Itti & Koch, 2000b)(Itti & Koch, 2001c) developed an algorithm to measure *visual saliency* (how likely people are to look at parts of an image) on the basis of image characteristics such as intensity distribution, color changes, and orientation. Saliency maps could prove to be a good candidate to indicate the initial attention in a painting. Then, by modifying the digital version of the painting to redistributef saliency, we could build several versions of the painting with the pre-selected interesting regions manipulated to increase saliency. For example, in "The Tribute Money" when it is time to look at Peter retrieving the coin from the mouth of the fish, SGD could boost the saliency in that region and thereby influence the viewer to re-direct their gaze.

This remainder of this paper presents a psychophysical experiment that explores the impact of SGD on performance during a viewing of narrative art works. The results show that this method works well without introducing noticeable artifacts into images, which might degrade the viewing experience.

3.3.4 Experimental Design

The goal of this experiment is to determine to what extent SGD lends itself to aid observers in extracting the intended sequence of events from regions of an episodic image. Participants viewed a sequence of images, each of which contained three or more panels intended as episodes. The intended viewing order of these panels is not always immediately clear. Art history research provides the narrative for each art piece, from which panels are determined (Velli, 2007). Panels (in each image) were manually selected as rectangular regions which enclosed the relevant portion of the image that conveyed an episode of the story. In the non-control group, after viewing the image for a short period of time, relevant panels were highlighted using SGD at the panel center. Participants then indicated the order they perceived to be the correct viewing order by clicking on image sections outlined with boxes. We compare performance using SGD with performance under normal viewing conditions.

Stimuli

Eleven images served as stimuli for the experiment, two of which were used for observer training, see Figure 3.22). In each image, episodes or *panels* were identified and served as targets for SGD. The number of episodes (panels) varied from painting to painting, ranging from three to seven.

The size of panels also varied within each image (see Figure 3), initially we were concerned that fixations may be artificially increased in proportion to panel size, but this did not turn out to be the case.

Presentation order was randomized to minimize the introduction of learning effects. Images were presented for a period of time proportional to the number of episodes in the image and were displayed on a 22 inch widescreen monitor, operating at 60 Hz with a resolution of 1680 x 1050. Stimuli transitioned directly from one to the next, however, no action was taken until a finite amount of time had passed. Image size varied as shown in Figure 3.22. In cases where images size was smaller than the viewing screen, width and height were maximized to fit screen resolution and a black border was added. An example of a single image, with all the regions highlighted for illustration purposes, is shown in Figure 3.18. In this image there are six panels. The observer would not see the regions highlighted in this obvious manner. This simply illustrates the presence of the panels.

Participants were seated in front of a computer screen in a well-lit room. Using a SensoMotoric Instruments iView X Remote Eye-Tracking Device operating at 250 Hz with gaze position accuracy $< 0.5°$, data

pertaining to fixation position and saccades were recorded for each participant. After a brief calibration phase, each observer underwent a short tutorial session to familiarize them with the experimental procedure and user interface. Questions were encouraged during the tutorial session but no data was collected. Participants were then presented with each of the nine art works in a random order. Image complexity varied, as did the number of panels. The two groups are as follows:

- **Group 1: Normal Viewing Conditions**: No actions were applied to the images, in other words images were viewed normally with no modulations. This group served as the control group for the experiment.
- **Group 2: Subtle Modulation:** SGD was employed to highlight the target panel regions in the intended viewing order in an effort to aid in visual navigation. Gaze manipulation was implemented as described in (Bailey et al., 2009b). Modulation was never applied to panels while they were being directly viewed. All modulations were only applied to the peripheral vision. Modulations were stopped as the observers gaze tended toward them i.e. observers never directly viewed modulations. A modulation radius of 0.04° of visual angle was used to ensure that modulations were subtle. The modulations were placed in the center of the panels.

Thirty-six participants were assigned randomly to one of the two groups. Participants were volunteers from a group of undergraduates. All had normal, or corrected-to-normal, vision and were naive to the purpose of the experiment. Viewing time for each image varied in direct proportion to the number of panels present. Each image was presented for t seconds before the user was allowed to respond. For the control group, Group 1, t was chosen to be equal to the number of regions in the image. In Group 2, t is the time taken to guide the viewers exactly once through the correct sequence of regions. Previous studies (Bailey et al., 2009b), (McNamara, Bailey, & Grimm, 2008b) revealed that SGD modulations typically attracted gaze within 0.5 seconds. To ensure that we had comparable viewing times between both groups a 0.5 second delay was added between successive modulations.

After t seconds, the relevant regions were highlighted with rectangles and the mouse activated to allow the users to respond. Both groups of participants were instructed to click on the highlighted regions in the order they believed the story unfolds. Each participant reported an order which they believed matched the intended sequence of the story in the art work.

Analysis of Data

In addition to recording eye-movements for each participants, each participant reported an order for each image, based on their understanding of panel sequence within that image. We needed a robust mechanism to compare accuracy of performance between the two groups. Levenshtein distance (Navarro, 2001) (V. Levenshtein, 1965) (V. I. Levenshtein, 1966) is a string metric, developed in the filed of information theory and computer science to compute *differences between sequences*. Levenshtein distance provides an appropriate measure to compare distances between ordered sequences, such as those recorded during our experiment. To accurately compare sequences using Levenshtein distance the correct (intended) viewing order of each image is converted into a string sequence. All responses from each participant are also converted to an appropriate string sequence in order to facilitate comparison to the correct sequence. Since the number of relevant regions varies across the images we normalize the distance measure computed for each image by dividing by the number of panels. The normalized Levenshtein distance L between the correct sequence $S_{correct}$ and user sequence S_{user} is as follows:

$$L = \frac{Levenshtein\ Distance(S_{correct}, S_{user})}{\#\ of\ Panels} * 100 \tag{3.3}$$

For example, let the correct panel order be [ABCDE] and let [ACBDE] denote the participant's response. The number of *panels* in this image is five. Using Equation 3.3, we obtain a normalized Levenshtein distance value of 40. A distance of 0 would indicate no difference, whereas a distance of 100 would indicate maximal distance.

Figure 3.18: An exemplar image showing all of the panels which contribute to the narrative. Observers first viewed the image without the panels highlighted. Once a certain amount of time had elapsed, participants then clicked on the panels in the order they believed matched the order of the story being told. The modulations for the SGD group (Group 2) were placed at the center of each panel.

3.3.5 Results and Discussion

The predicted sequence of panels reported by each participant for each image was recorded. Normalized distances for each image were compared to the actual intended sequence for that image using the distance metric expressed in Equation 1. The calculated normalized Levenshtein distance measures between conditions showed differences across the two groups with a mean distance measures of 57.32 and 34.79 for groups 1 and 2 respectively, as illustrated in Figure 3.19. These values were calculated by averaging the normalized Leveshtein distance (L) for all the participants in a group over all the images used in the experiment. This implies that participants from Group 2 (guided by SGD) consistently proved to be more accurate at predicting the intended sequence of panels contained in narrative art when compared to Group 1, the control group (no SGD). This measure indicates that, for example, when viewing a narrative art image containing ten panels, the static viewing group will incorrectly predict the order of approximately $5-6$ panels, while the gaze directed group, Group 2, will return a prediction with only $3-4$ panels out of the sequence. An independent-samples paired $t-test$ suggests that this was a significant effect:

$$t(316) = 1.9675; p < 0.05 \tag{3.4}$$

Figure 3.20 shows the average L value for each image across all the participants in each group. This analysis gives some intuition on the the influence of the number of panels over the accuracy in detecting the correct sequences in the narrative art. The images in the graph are arranged in the increasing order of number of panels. In all nine images, the average L value indicates that the participants in Group 2, the gaze directed group, predict panel order more accurately than the participants from Group 1, the static viewing group. This also shows that the gaze directed group performed better than the static group for images having relevant regions varying from 3 to 26. In eight of the nine images used in the study, this result was shown to be significant. Independent-samples t-tests reveal that this effect was significant and not due to chance, the t-test results for images with relevant regions 3,4,5,6,7 and 26 are shown in Table 3.2.

The results did reveal a single anomaly where the t-test did not show a significant difference between

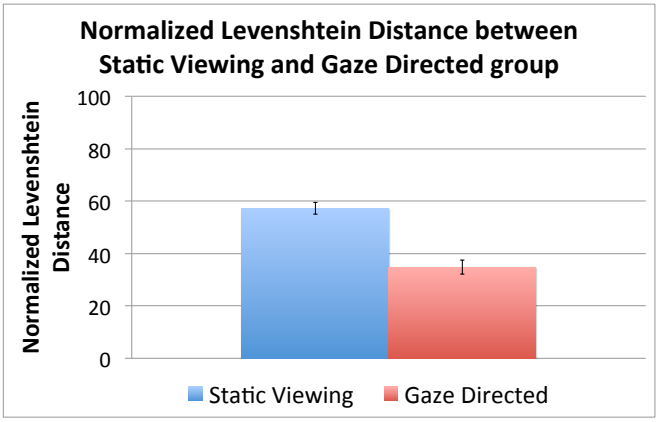


Figure 3.19: Normalized Levenshtein distance measure between Group 1 (static viewing group) and Group 2(gaze directed group). Error bars represent one standard error. The graph shows that Group 2 participants, who viewed SGD images were able to predict the intended viewing order of panels more accurately than those in Group 1 that did not have the benefit of SGD as a gaze direction aid.

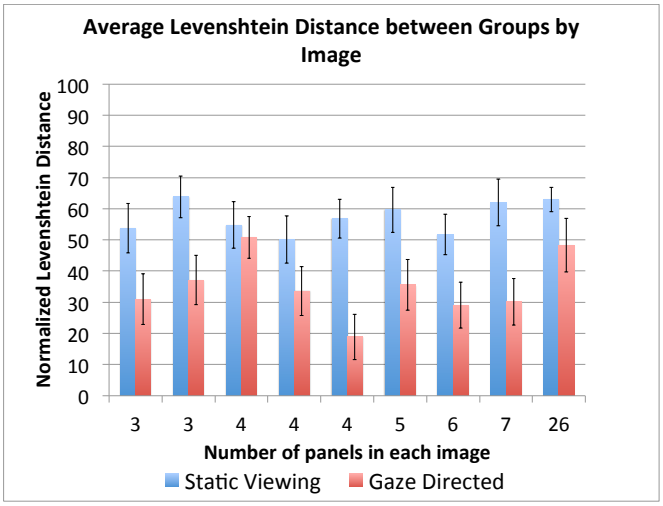


Figure 3.20: Normalized Levenshtein distance measure between static viewing group and gaze directed group for each image. The x-axis indicates the number of panels in each image. The error bars represent one standard error.

groups. Image *"C"*, shown in Figure 3.22, revealed no significant difference in performance between the two groups. Further inspection showed that in this image, the artist has gradually decreased the luminance of the narrative art over the story. This analysis was possible as the same characters appear over multiple regions in the image. We reason that this luminance change in itself would provide a strong enough visual cue to enable the participants in Group 1 to correctly navigate the story. This also suggest that luminance changes could serve to guide direct gaze in imagery. This phenomenon is a topic for future research.

To further illustrate the success of SGD we present a single example. Figure 3.21 shows the images for the static viewing and the gaze directed group respectively. Images are placed side-by-side for comparison.

Image A depicting the scan path of the viewer's gaze over the static images shows that the viewer's gaze does not coincide with all of the relevant story panels. Contrast this with image B, which shows the scan path over the SGD enhanced image. This image reveals a more coherent scan path in terms of visitation to each relevant panel.

Comparing heat maps reveals a similar story. The heat maps represent the amount of time spent fixating in each image region. Figure 3.21,Image C reveals that most fixations fall to the left of the image, and the

Number of panels	Independent t-test (Group 1 V Group 2)
3	t(33) = 2.0322
4	t(33) = 2.0346
5	t(33) = 2.0345
6	t(33) = 2.0340
7	t(33) = 2.028
26	t(33) = 2.0364

Table 3.2: Independent t-tests indicate significant differences in the ability to correctly predict intended panel sequences for images with vary in numbers of panels. In each case, $p < 0.05$.

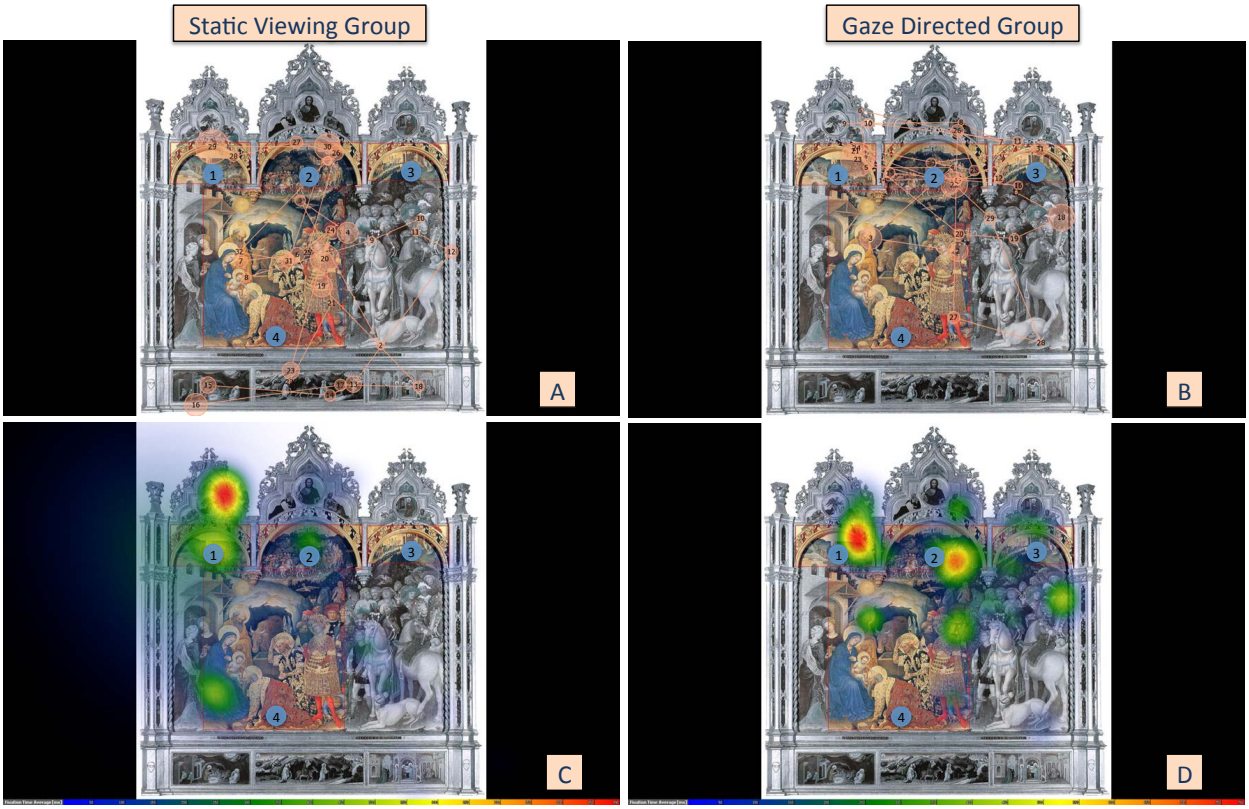


Figure 3.21: This Figure shows the scan paths and heat-maps for one participant from each of the two groups (no modulation and SGD), Image A & C represent data collected from Group 1, while images B & D were collected from Group 2. Rectangular highlighted regions denote panels. Blue numbered circles indicate the correct viewing order of panels within the image. As can be seen from the images, gaze distribution and fixations are more accurately aligned with panel (modulated) regions in the SGD condition.

distribution does not encompass the story panels. Conversely, examination of the heat map for image D (SGD) indicates that viewer fixations are distributed over the story panels.

Examining the L value measures (as described in Section 3.3.4) for this single image, the Group 1 (no SGD) participant's value is 71.45 compared to 28.57 for the participant from Group 2 (the gaze directed group). Thus the heat map and scan path analysis not only reflect the increase in the gaze coverage and attention to all relevant regions of the image for the gaze directed group over the static viewing group, but also correspond well with the L-value metric chosen to compare performance.

It is important to note that Figure 3.21 serves as a representative example of a consistent trend across all nine images viewed. This analysis reveals that, *without SGD*, not only did participants fail to view all

of the story panels, but they failed to fixate on all the relevant story panels. The exact opposite is true for those images presented with SGD applied to the story panels, giving a high level of confidence in the success of applying SGD to subtly reveal an intended viewing sequence.

Informal reporting, post experiment, showed that a small subset of participants (approximately 15%) reported noticing the modulations, but (as designed) the modulation disappeared before they could inspect it. A single participant reported trying to follow the modulations, again performing exactly the action SGD is designed for. In these cases we reason that certain individuals may have heightened peripheral vision e.g. it has been proposed by several researchers, for example, that strong peripheral vision may give basketball players, and other athletes a distinct advantage during game play (Vickerss, 2007). In general, however the premise of subtly holds well for SGD. Even in cases where individual participants noticed SGD task performance was not impeded.

In summary our results show that gains can be made in task performance when modulation is employed to direct gaze to target a sequence of panels in a specific order. Even if the modulations are noticed to some degree, there is still an increased accuracy in task performance. This seems to hold true over a range of images, and over a range of panel numbers. For applications that require a specific viewing order for understanding or performance, SGD can serve as a subtle aid to boost accuracy of performance on a sub-image ordering task.

3.4 Conclusions and Future Work

We presented an experiment to compare task performance in digital images across two groups of stimuli. In one group no image alterations were used (Group 1), in the second group small modulations were applied to image panels in an effort to direct the viewers gaze (Group 2). The participant's task was to specify the order of panels (contained in episodic art pieces) which revealed the intended story. The results indicate that using a subtle gaze direction technique, which modulates the appropriate panel in the intended sequence, does indeed improve the precision of panel ordering. The difference between performance between the two groups was shown to be significant.

For this study we chose to modulate the luminance channel within a small radius of pixels. All modulations took the same shape and were modulated using the same oscillation strength. Given the variation of the subset of images we included in this study, and the variation in the number of panels, it may be that variation of modulation shape and strength would yield further improvements in task performance. For example, when the number of relevant panels is large, perhaps a stronger modulation would attract gaze faster. For future work it may be useful to implement modulations characteristics based on image content rather than take a "one-size-fits-all" approach.

Also, in this chosen domain, all of the imagery is static. We are interested in applying SGD to dynamic environments, which may pose increased difficulty due to the attentional draw of objects in motion. We anticipate that stronger modulations will be required to successfully used SGD effectively in dynamic environments, but the payoff could be beneficial in areas such as simulation training and educational gaming.

This section has shown how SGD can improve performance on a within-image panel ordering set without noticeably disrupting the visual experience of the image. This technique can be applied to help guide gaze in complex applications where viewing order is critical to understanding, such as in story telling, or training task performance where sequence of operations is important, for example, construction instructions.

Figure 3.22: Narrative artwork images used in this study, from top left to right, The Miraculous Draught of the Fishes (Konrad Witz) (Witz, 1434), Adoration of the Magi (Gentile da Fabriano) (da Fabriano, 1423), The Tribute Money (Masaccio), (Masaccio, 1421), Christ Taken Prisoner (Duccio di Buoninsegna) (di Buoninsegna, 1311a), The Story of Joseph (Biagio d'Antonio), (dAntonio, 1485) reproduced with kind permission of The J. Paul Getty Museum, Los Angeles, Landscape with Perseus and Andromeda, (Boscotrecase, 1BC), Maesta Altarpiece (Duccio di Buoninsegna), (di Buoninsegna, 1311b).

3.5 Gaze Direction in Medical Applications

This section has been adapted from Sridharan et al. (Sridharan, Bailey, McNamara, & Grimm, 2011).

3.5.1 Introduction

The American Cancer Society estimates that 250,000 new cases of breast cancer were diagnosed in the US in 2009-2010 (American Cancer Society, 2014). The statistics also show that breast cancer is the leading form of cancer among women. Early diagnosis and treatment of breast cancer are essential to maintaining the patients long-term health. In addition to breast self-exams and clinical breast exams, mammograms play an important role in the early detection of breast cancer. Radiologists undergo extensive training to become proficient at reading mammograms. In the US, an expert radiologist typically completes four years of undergraduate study, four years of medical school, one year of internship, four years of residency training and one or two years of additional fellowship training. Despite advances in technology, radiological training still uses the conventional approach of having a trainee work alongside an expert radiologist.

In this section we present a novel training technique that uses gaze manipulation to guide novices along the recorded scanpath of an expert radiologist. We accomplish this using the Subtle Gaze Direction technique (SGD) (Bailey, McNamara, Sudarsanam, & Grimm, 2009c) which combines real-time eye-tracking with

Figure 3.23: Photograph of experiment setup. The eye-tracking hardware is fixed to the bottom of the screen. Participants were asked to identify irregular regions in a sequence of mammogram images. Gaze manipulation techniques were used to influence their sequence of fixations during the experiment.

subtle image-space modulation. It has minimal impact on the viewing experience as it does not change the overall appearance of the image being viewed. Subtlety is achieved by presenting the modulations only to the low-acuity peripheral regions of the field of view so that the viewer is never allowed to scrutinize the modulations. The technique has been shown to be quite fast and accurate: viewers typically attend to target regions within 0.5 seconds of the onset of the modulation and the resulting fixations are typically within a single perceptual span of the target. While this shows that the technique is successful in directing gaze, its usefulness for training applications has not yet been established. This paper explores the application of Subtle Gaze Direction within the context of digital mammography training.

This work hypothesized that guiding a novice along the scanpath of an expert radiologist would improve the likelihood that the novice correctly identifies the abnormalities in the mammograms. To test this hypothesis, we designed an experiment to explore whether Subtle Gaze Direction is capable of improving the efficiency of digital mammography training. During a training session we noticed that actively guiding participants along the expert's scanpath or directing them to the regions marked by the expert significantly improved their accuracy compared to a control group who viewed the mammograms without gaze manipulation. We also conducted two follow-up sessions without the use of gaze manipulation, a short-term follow-up session (5 minutes later) and a long-term follow-up session (1 week later), to determine if the novice participants had become sensitized to the strategies used during the training session. In the short-term follow-up session, we observed that the groups of participants who were guided using SGD continued to perform better than the control group. This effect was not observed in the long-term follow-up session.

3.5.2 Background

Understanding the diagnostic process used by radiologists when searching for cancerous regions in mammograms is an interesting and active area of research. Studies have been conducted to analyze the fixation pattern of expert radiologists when viewing mammogram images. Kundel et al.(Kundel, Nodine, Krupinski, & Mello-Thoms, 2008) recorded and analyzed the fixations of expert radiologists from three independent institutions as they studied cancerous regions on mammograms. Their work provides useful information about the duration of expert viewing and identifies several regions of interest in digital mammograms. Mello-Thoms et al. (Mello-Thoms et al., 2006) also analyzed dwell time and number of fixations by experts during scans for breast cancer in mammograms using both head mounted and remote eye tracking devices. Krupinski et al. (Krupinski, 1996) studied decision making among mammographers and radiology residents with gaze duration as the key parameter.

The study of expert gaze patterns within the medical field is not limited only to breast cancer research and mammograms. Krupinski (Krupinski, 2000) summarizes work in this area and discusses the importance of perception research in medical imaging. Carmody et al. compare fixations between instructors and radiology residents for lung scan images (Carmody, Nodine, & Kundel, 1980) and chest scans (Carmody, Kundel, &

Toto, 1984). Sowden et al. (Sowden, Davies, & Roling, 2000) used eye tracking to study how perceptual learning affects a subject's ability to detect features in general X-ray images. Litchfield et al. (Litchfield, Ball, Donovan, Manning, & Crawford, 2010) (Litchfield & Ball, 2011) showed that viewing an expert's eye movements can help to improve identification of pulmonary nodules in chest x-rays and aids problem solving.

Image processing and computer vision techniques can be used to automate the process of abnormality detection in mammograms and other medical images. Neural network techniques for example, have been used to detect solid breast tumors on ultrasonic images (Chen, Chang, & Huang, 1999). Other techniques use first- and second-order histogram statistics on segmented regions and local textures to diagnose micro-calcifications in malignant breast tumors (Dhawan, Chitre, & Kaiser-Bonasso, 1996). Markov's Random Field segmentation can be used to detect suspicious regions and distinguish them from normal regions using a fuzzy binary decision tree (H. Li, Kallergi, Clarke, Jain, & Clark, 1995). While these techniques automate the process of abnormality extraction in medical images, it is still necessary for the results to be validated by an expert radiologist.

Currently most radiological training programs still use the conventional approach of having a trainee work alongside an expert radiologist. We propose a novel computer-based training technique that uses gaze manipulation to guide novices along the recorded scanpath of an expert radiologist. Computer-based training systems and workstations have become more popular in the medical field with the emergence of digital medical images and improved computer graphics and visualization techniques (Gay, Sobel, Young, & Dwyer, 1997; Dugas et al., 2001; Sharples et al., 2000; Muniyandi, Cotin, Srinivasan, & Dawson, 2003; Desser, 2007; Chentanez et al., 2009; Relan et al., 2009).

Strategies for enhancing performance on visual search tasks range from guiding attention to previously unattended regions (Qvarfordt, Biehl, Golovchinsky, & Dunningan, 2010) to guiding attention directly to the relevant regions in a scene (Grant & Spivey, 2003; Thomas & Lleras, 2007; R. F. Wang & Spelke, 2002). Such techniques have been shown to improve reaction time and enhance problem solving capabilities. In the experiment presented in this paper, we recruit subjects having no prior knowledge about mammogram images and use the Subtle Gaze Direction technique to guide them along the scanpath of an experienced radiologist as they search for irregularities. In our experiment we also analyze the post-training effects on the subjects guided using Subtle Gaze Direction compared to the subjects in the control group who were not gaze directed.

3.6 Experiment Design

This section describes an experiment conducted to investigate if subtle gaze manipulation is capable of improving the efficiency of digital mammography training. During a training session, participants viewed a randomized sequence of mammogram images and were asked to identify what they considered to be irregularities. There were four groups of participants. One group viewed the images without the use of gaze manipulation. They served as the control group for the experiment. For the other groups, the Subtle Gaze Direction technique was used to guide the participants in various ways. Two follow-up sessions were completed without gaze manipulation to determine if the participants became sensitized to the training method used.

Stimuli

The database of mammogram images used in this study was provided by the Mammographic Image Analysis Society (Suckling et al., 1994). The database contains pairs of mediolateral images from 161 patients along with a separate text file containing information collected from an expert radiologist such as the x,y image coordinates of the center of abnormalities and the approximate radius in pixels of a circle enclosing the abnormality. What is missing however, is the expert's scanpath when trying to locate these abnormalities. We hired our own expert radiologist to view a subset of 65 images from the database and to mark any abnormalities present by drawing a circle enclosing the abnormality. The expert's scanpath was recorded during this process and later used to guide one group of the novice participants.

Stimuli for the experiment were presented on a 22 inch widescreen monitor, operating at 60 Hz with a resolution of 1680 x 1050. The stimuli consisted of the 65 pairs of mammogram images that were viewed by the expert. Five of these images were used for an initial tutorial session to familiarize the participants

with the software interface. Twenty images were used for a training session, twenty images were used for a short-term follow-up session, and twenty images were used for a long-term follow-up session. For all sessions, the participants were instructed to identify what they considered to be irregularities in the images. For all sessions, images were displayed for 10 seconds before input from the user was accepted. For the gaze-directed groups, the Subtle Gaze Direction technique was applied during this time frame (but not during the subsequent marking stage). This helped to minimize the impact of viewing duration on the results and also allowed enough time to guide the viewer's gaze about the images. Ten seconds was chosen based on the expert's average viewing time before making the first selection (9.71 seconds).

The pairs of images for each patient were arranged so that the image of the left breast was placed on the right and the image of the right breast was placed on the left to mimic the preferred arrangement of the expert. Figure 3.23 shows a participant viewing one of these images on a monitor connected to an eye-tracking device. The eye-tracker used in this study is a SensoMotoric Instruments iView X Remote Eye-Tracking Device operating at 250 Hz with gaze position accuracy $< 0.5°$.

Participants

20 novice participants (7 females, 13 males), between the ages of 18 and 30 volunteered to participate in this study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of four groups:

- **Static group:** 5 participants were presented with a randomized sequence of the 20 training images without the use of gaze manipulation. This group served as the control group for the experiment.
- **Gaze-directed group using expert scanpath :** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them to follow a similar scanpath (sequence of fixations) as the expert.
- **Gaze-directed group using expert selections :** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them only to the regions marked by the expert as irregularities. The overall scanpath of the expert was not used.
- **Gaze-directed group using adversarial scanpath:** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them along a scanpath that was chosen by the researchers to follow different directions and locations than that of the expert.

We hypothesized that using gaze manipulation to guide a novice along a similar scanpath as the expert or directly to the locations marked by the expert would improve the likelihood in correctly identifying irregular regions on the mammograms. Similarly, we hypothesized that using gaze manipulation to guide the novice's focus along an adversarial scanpath would reduce the likelihood that the novice correctly identifies the irregular regions on the mammograms.

Procedure

The expert radiologist hired to participate in this study helped to guide the user interface design of our training platform so that it closely mimics that of standard digital mammography systems. The expert also described external factors such as preferred lighting conditions for studying mammograms which we incorporated into the design of our experiment. The experiment was divided into four stages — a tutorial session, a training session, and two follow-up sessions to determine if the novice users had become sensitized to the training method used.

Tutorial Session

All participants took part in the tutorial session. They were each presented with the same five images in a randomized order and asked to identify what they considered to be irregularities in the images. The images were displayed for 10 seconds before user input was accepted. A left mouse click-and-drag motion was used to draw a circular region that enclosed the abnormality. The Shift key was locked in software to ensure that

the region drawn was always circular. Multiple regions could be selected in this manner. A double-click moved on to the next image. The sole purpose of the tutorial session was to get users familiar with the user interface and method of region selection to be used in the remaining stages of the experiment. No feedback about the accuracy of their selected regions was given.

Training Session

All participants took part in the training session. They were presented with the same twenty images in a randomized order and asked to identify what they considered to be irregularities in the images. For participants in the gaze-directed groups, the Subtle Gaze Direction technique was used to influence the order and location of their fixations during the fist 10 seconds of viewing. Subtle Gaze Direction was implemented as described in (Bailey et al., 2009c), i.e. viewer gaze was monitored in real time to ensure that modulations were only presented to the peripheral regions of the field of view and modulations were immediately terminated as the viewer's focus approached the modulated regions. We found that no single modulation intensity worked well in all cases because of the wide variation in mammogram images from black to white. Some required more intense modulation than others to attract attention in the peripheral vision. To overcome this problem, the intensity was set to increase gradually until it captured the user's attention as evidenced by a saccade toward the modulated region.

To determine the location and order of the modulations for the expert-scanpath gaze-directed group, the expert's scanpath was replayed in slow motion and the researchers manually selected what they considered to be the center of each cluster of fixations in sequence. This manual approach worked well since there were only 20 expert videos to consider. Of course, clustering algorithms (R. Xu & Wunsch, 2005) or connected components analysis (Gonzalez & Woods, 2006) could be used to automate this process. Using these simplified expert scanpaths as a guide, the researchers also manually selected the order and locations of the modulations for the adversarial scanpaths. These were chosen to follow different directions and locations than that of the expert. Figure 3.24 shows an example of the expert's raw scanpath and the resulting simplified scanpath and adversarial scanpath. It also shows an overlay of both the simplified scanpath and the adversarial scanpath to illustrate how different they are. For the expert-selection gaze-directed group, the modulations corresponded to the centres of the regions marked by the expert as containing irregularities. The modulations were presented in the order they were marked.

Short-Term Follow-up Session

Following the training session, the participants were given a five minute break. A follow-up session was run using twenty fresh images. No gaze manipulation was used in this session for any of the participants. The purpose of this follow-up session was to determine if the novice participants had become sensitized (short-term) to the strategies used during the training session.

Long-Term Follow-up Session

The participants were asked to return one week later for a second follow-up study. Similar to the first follow-up session, this session was run using twenty fresh images and no gaze manipulation was used for any of the participants. The purpose of this follow-up session was to determine if the novice participants had become sensitized (long-term) to the strategies used during the training session.

3.6.1 Results

To facilitate analysis of the novices' performance, we define a weight-based accuracy measure as follows:

$$Accuracy = \left(\frac{1 + (w_h * h + w_c * c + w_m * m)}{1 + h + c + m} \right) * 100 \tag{3.5}$$

where h is the number of hits, c is the number of close matches, and m is the number of misses. w_h, w_c, and w_m are corresponding weights. Figure 3.25 illustrates the concepts of hits, close matches, and misses which are defined in terms of the circles drawn by the expert and the circles drawn by the novice. A *hit* occurs if

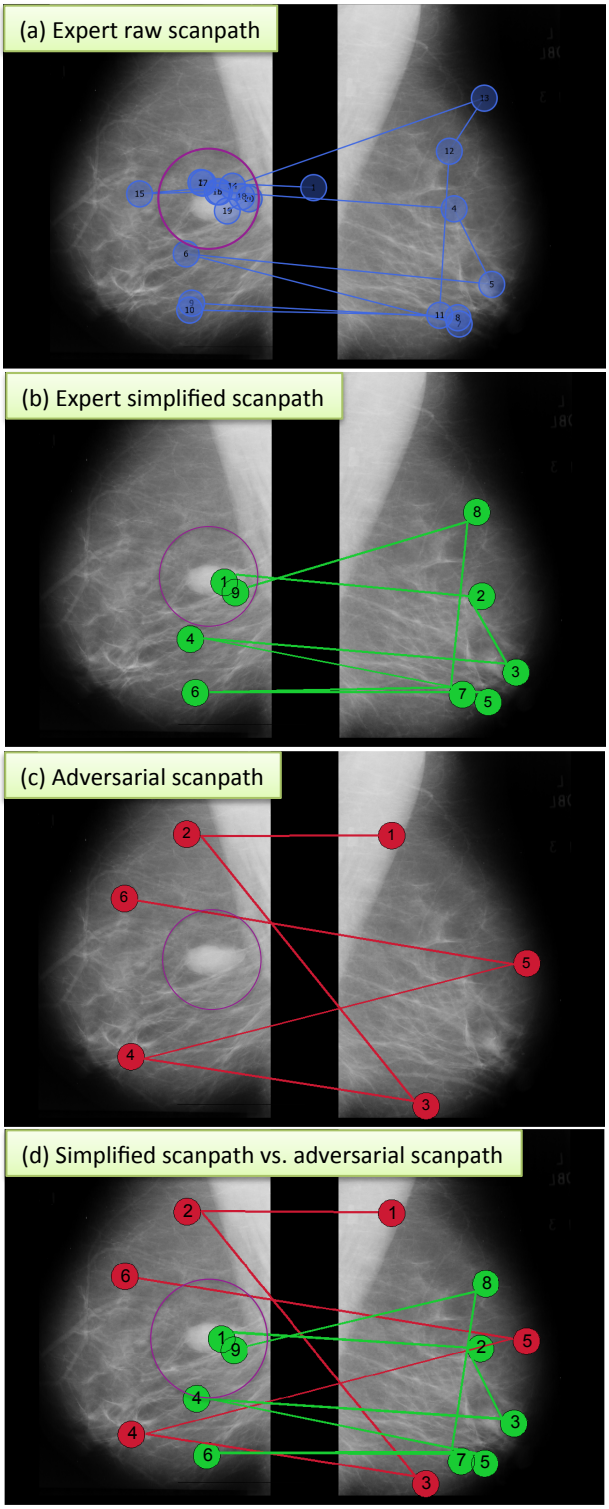


Figure 3.24: Expert's raw scanpath (a), simplified expert scanpath (b), adversarial scanpath (c), and comparison of simplified expert scanpath and adversarial scanpath (d) for one image from the dataset. The large purple circle shows the area marked by the expert as an irregularity. The numbers in the smaller circles of (a) indicate the order of fixations of the expert. The numbers in the smaller circles of (b) and (c) indicate the order of modulations used to guide participants in the corresponding groups.

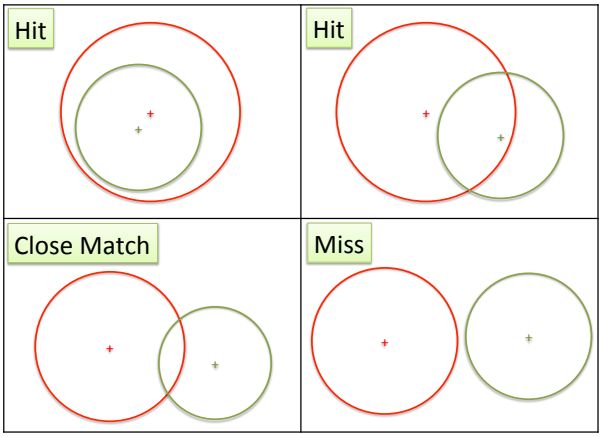


Figure 3.25: Illustration of hits, close matches, and misses. The circles represent selections by the expert and novice. The center of the circles are represented by the crosses.

the distance between the centers is less than either of the radii as shown in the top row of Figure 3.25. A *close match* occurs if the distance between the centers of the circles is less than the sum of the radii as shown in the bottom left of Figure 3.25. Finally, a *miss* occurs if the circles do not overlap as shown in the bottom right of Figure 3.25. If multiple novice selections result in hits to a single region selected by the expert, only one is considered to be a hit and the others are ignored. Likewise if multiple novice selections result in close matches to a region selected by the expert, only one is considered to be a close match and the others are ignored. In the case where there is a mix of hits and close matches to a single region selected by the expert, only one hit is considered and the others are ignored. We assign the following weights: $w_h = 1$, $w_c = 0.75$, and $w_m = 0$. Note that the weight for a close match is biased more towards a hit than a miss. The reason for this is that the selection is sufficiently close to an abnormality to warranty further investigation instead of labeling it simply as a random selection.

Alternatively, binary classification statistics (Agresti, 1989) (Fawcett, 2004) can be used to establish measures of accuracy as well as sensitivity and specificity. To calculate these properties it is necessary to categorize the test outcomes as *true positives*, *true negatives*, *false positives*, and *false negatives*. Table 3.3 shows the mapping of test outcomes to these conditions.

True positive (TP)	Abnormal regions exist, and are correctly identified by the subject
False positive (FP)	No abnormal regions exist, yet subject marked regions as being abnormal
True negative (TN)	No abnormal regions exist, and no regions were marked by subject as being abnormal
False negative (FN)	Abnormal regions exist, but were not marked by the subject

Table 3.3: Mapping of test outcomes to true positives, true negatives, false positives, and false negatives.

For this study *sensitivity* refers to the ability of the subject to correctly identify existing abnormalities in the mammograms. Sensitivity is computed as follows:

$$Sensitivity = \frac{(\#ofTP)}{(\#ofTP + \#ofFN)} * 100 \tag{3.6}$$

On the other hand, *Specificity* refers to the ability of the subject to correctly identify mammograms containing no abnormalities. Specificity is defined as follows:

$$Specificity = \frac{(\#ofTN)}{(\#ofTN + \#ofFP)} * 100 \tag{3.7}$$

The sensitivity and specificity values can then be combined to produce a binary classification based measure of accuracy as follows:

$$Accuracy = \frac{(\#ofTP + \#ofTN)}{(\#ofTP + \#ofTN + \#ofFN + \#ofFP)} * 100 \tag{3.8}$$

Training Session Results

Figure 3.26 shows the average weight-based accuracy (equation 3.5) for the various groups of participants during the training session. The static group averaged 54.2%, the group that was guided by the expert scanpath averaged 64.5%, the group that was guided by the expert selections averaged 68.8%, and the group that was guided by the adversarial scanpath averaged 55.5%. These values were obtained by averaging the accuracy for all participants in a group over all images used in the training session.

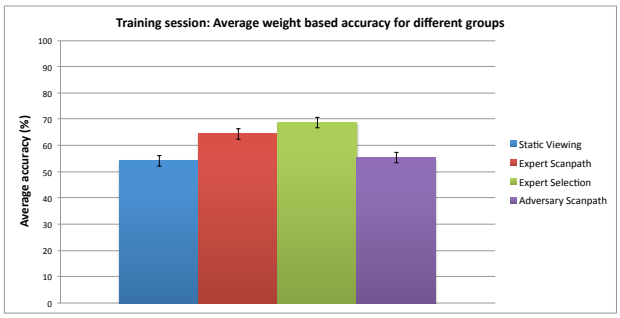


Figure 3.26: Average weight-based accuracy (equation 3.5) for different groups during the training session. The error bars represent standard error.

The averages show that the participants from both expert guided groups were more accurate than the participants from the static group. Independent-samples t-tests reveal that this effect was significant in both cases and not due to chance:

$$t(198) = 3.227; p < 0.05 \ \ (expert\ scanpath\ vs.\ static)$$

$$t(198) = 4.530; p < 0.05 \ \ (expert\ selection\ vs.\ static)$$

Figure 3.27 shows the binary classification based accuracy (equation 3.8) for the various groups of participants during the training session. The static group averaged 53.6%, the group that was guided by the expert scanpath averaged 68.3%, the group that was guided by the expert selections averaged 64.7%, and the group that was guided by the adversarial scanpath averaged 52.0%. When using this accuracy measure we again observe that both expert-guided groups perform better than the static group.

These observations are not surprising as previous studies have already established that guiding attention to the relevant regions of a scene facilitates task completion (see section 3.5.2). With Subtle Gaze Direction however, there is the added benefit that the cues used to attract the viewer's attention have minimal impact on the viewing experience as they occur only in the viewer's peripheral vision and do not permanently alter the overall appearance of the image being viewed.

Short-Term Follow-up Session Results

Figure 3.28 shows the average weight-based accuracy (equation 3.5) for the various groups of participants during the short-term follow-up session. The static group averaged 51.3%, the group that was guided by

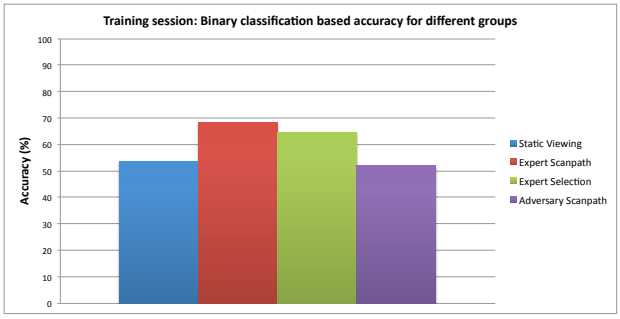


Figure 3.27: Binary classification based accuracy (equation 3.8) for different groups during the training session.

the expert scanpath during the training session averaged 59.5%, the group that was guided by the expert selections during the training session averaged 63.1%, and the group that was guided by the adversarial scanpath during the training session averaged 53.9%. These values were obtained by averaging the accuracy for all participants in a group over all images used in the short-term follow-up session.

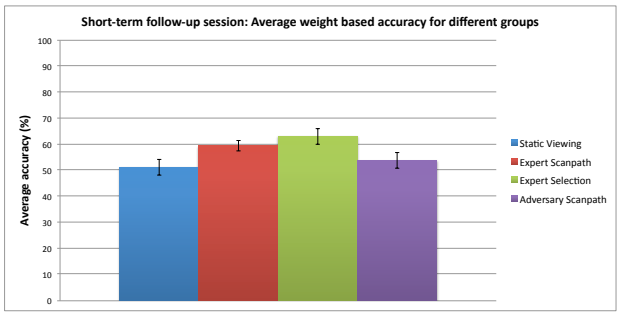


Figure 3.28: Average weight-based accuracy (equation 3.5) for different groups during the short-term follow-up session. The error bars represent standard error.

The averages show that the participants who were trained using the expert-guided approaches performed better than participants who were not gaze directed during training. This indicates that there are short-term lingering effects related to the use of gaze manipulation and that the participants are becoming sensitized to the method of training used. Independent-samples t-tests confirm that this effect was significant in both cases and not due to chance:

$$t(198) = 2.311; p < 0.05 \ \ (\textit{expert scanpath vs. static})$$

$$t(198) = 3.189; p < 0.05 \ \ (\textit{expert selection vs. static})$$

Figure 3.29 shows the binary classification based accuracy (equation 3.8) for the various groups of participants during the short-term follow-up session.The static group averaged 26.9%, the group that was guided by the expert scanpath averaged 45.8%, the group that was guided by the expert selections averaged 41.6%, and the group that was guided by the adversarial scanpath averaged 27.5%. When using this accuracy measure we again observe that both expert-guided groups continued to perform better than the static group in the short-term follow-up session.

Long-Term Follow-up Session Results

Figure 3.30 shows the average weight-based accuracy (equation 3.5) for the various groups of participants during the long-term follow-up session. The static group averaged 64.2%, the expert scanpath group averaged 64.2%, the expert selection group averaged 65.4%, and the adversarial scanpath group averaged 65.2%.

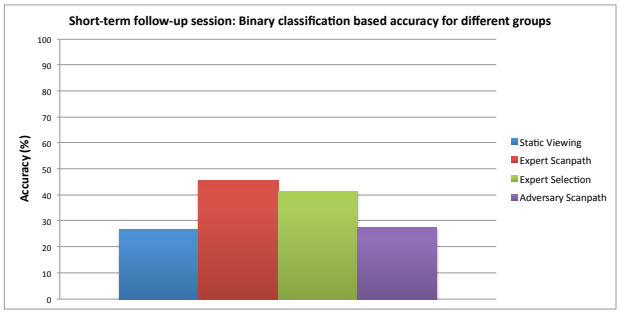


Figure 3.29: Binary classification based accuracy (equation 3.8) for different groups during the short-term follow-up session

These values were obtained by averaging the accuracy for all participants in a group over all images used in the long-term follow-up session. The averages show that there is little difference between the groups and independent-samples t-tests confirm that the differences are not significant. This suggests that there are no long-term lingering effects related to the use of gaze manipulation.

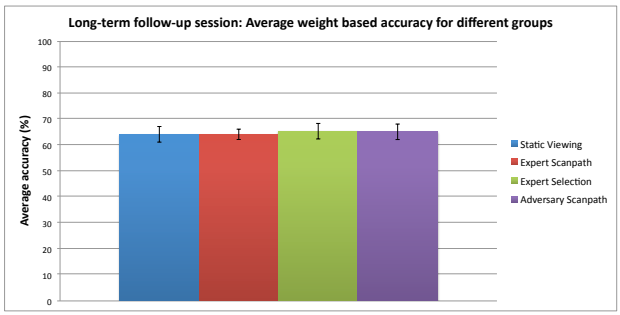


Figure 3.30: Average weight-based accuracy (equation 3.5) for different groups during the long-term follow-up session. The error bars represent standard error.

Figure 3.31 shows the binary classification based accuracy (equation 3.8) for the various groups of participants during the long-term follow-up session. The static group averaged 51.6%, the group that was guided by the expert scanpath averaged 59.1%, the group that was guided by the expert selections averaged 53.0%, and the group that was guided by the adversarial scanpath averaged 51.5%. When using this accuracy measure we again observe that there is little variation between the groups with the exception of the expert scanpath group which appears to be higher. Further analysis using the receiver operating characteristics (ROC) space (described below) shows however, that all of the groups are actually performing at the accuracy level equivalent to a random guess.

Receiver Operating Characteristic (ROC) Results

Figure 3.32 shows the sensitivity and specificity for each group of participants across all sessions. Using these values we can generate *receiver operating characteristic (ROC)* plots for each session as shown in Figure 3.33. ROC plots (i.e. *sensitivity* vs. *1-specificity*) help us assess the performance of the training techniques used. The blue dotted line at 45 ° represents the line of random guess. Any data above this line indicates a better-than-random classification and below this line a worse-than-random classification.

The ROC plot for the training session shows that the group guided by expert selection had a higher specificity compared to the other groups. The higher specificity may be attributed to the fact that the subjects associated the presence of the subtle modulations with the existence of abnormalities. Consequently when no modulations were present they were less likely to mark the image as being irregular. The ROC plot for the training session also shows that both expert-guided groups were more sensitive than the static and

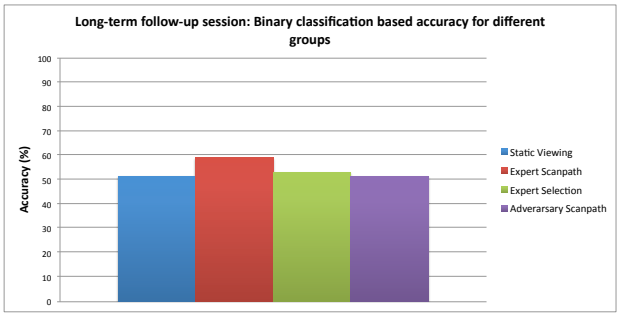


Figure 3.31: Binary classification based accuracy (equation 3.8) for different groups during the long-term follow-up session

	Static Viewing		Expert Scanpath		Expert Selection		Adversarial	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	63.19	21.15	72.92	30.30	66.31	59.26	64.12	27.66
Short term	50.15	7.07	63.13	19.23	46.13	33.33	46.75	11.65
Long term	53.33	47.06	68.87	33.33	56.21	45.71	55.56	41.18

Figure 3.32: Sensitivity and specificity results for training, short-term follow-up, and long-term follow-up sessions.

adversarial groups. This supports the accuracy findings described above.

The ROC plot for the short-term follow-up session shows that the expert-guided groups perform better than the control group and the adversary group. Note however that both the sensitivity and specificity for all groups is lower than in the training session. This suggests that there was some common phenomenon affecting the subjects of all groups. We believe that this could be due to fatigue or boredom (recall that the participants had already viewed 25 images lasting a minimum of 10 seconds each - not including the time it took for them to respond to each image).

The ROC plot for the long-term follow-up session shows that the performance of all the groups lie on the random guess line indicating that there are no difference between the groups and that there are no long-term lingering effects associated with the training methods used. Notice also that the phenomenon that caused the lower sensitivity and specificity observed during the short-term follow-up session is no longer present after a week-long break.

3.6.2 Summary and Conclusion

In our experiment we explored whether subtle gaze manipulation is capable of improving the efficiency of mammography training. Using a weight-based accuracy measure we observed the following:

- For the training session, participants in both expert-guided groups performed significantly better than the participants in the static group.
- For the short-term follow-up session, participants in both expert-guided groups continued to perform significantly better than the participants in the static group.
- For the long-term follow-up session, no significant effects were observed.

Similar results were observed using the binary classification accuracy measure for all sessions. The general agreement between these two accuracy measures provides added assurance about the reliability of the results.

Based on the ROC plots we can conclude that actively guiding the subjects using the expert-based approaches results in better performance in terms of sensitivity and specificity. In the short-term follow-up session we noticed that both sensitivity and specificity of the group guided by the expert selections dropped dramatically compared to the group guided using the expert scanpath. For this reason, and for the fact that

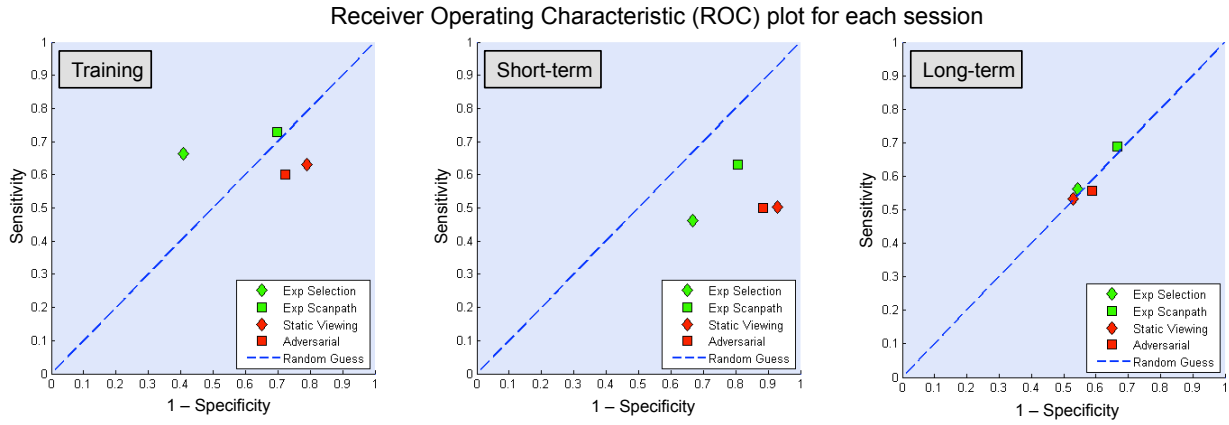


Figure 3.33: ROC plots for each session of the experiment for all groups. From left to right: training session, short-term follow-up session, and long-term follow-up session.

constantly modulating a single selection may become annoying, we conclude that it is better to guide the novices using the expert scanpath.

Overall, our findings indicate that using the Subtle Gaze Direction technique to actively guide novice participants along the scanpath of an expert significantly improves accuracy compared to static viewing. In addition to exhibiting great promise for digital mammography training as well as training on other medical image modalities such as X-rays, CT scans, PET scans, and MRIs, these observations have implications for a wide range of visual search and learning applications including:

- **Surgical Simulator Training:** Surgical simulation systems are becoming more common. Many of these systems are complex and require significant time and attention from the trainees. Gaze manipulation may facilitate more efficient training on these platforms.
- **Education and Learning:** Often the amount of information presented to students in educational settings, be it in the classroom or online, is overwhelming. Gaze manipulation can potentially be used to enhance the learning experience and benefit those who have difficulty paying attention during learning.
- **Aiding Symbolic Development in Children:** Children with early childhood developmental disorders often have difficulty associating words to images and symbols. Subtle Gaze Direction can be used in a symbolic teaching system to assist these children in leaning new symbols and images and associate them with the words they are hearing. The benefit of this approach is that it does not introduce any permanent visual cues which may complicate the learning process.

In addition to these applications, there are several other avenues for future work. We would like to explore if prolonged exposure (i.e. multiple training sessions using SGD as opposed to one) would have any long-term impact on the novices' performance. Further analysis of the eye-tracking data will be conducted to quantify how much deviation was present between the scanpaths of the novices and the expert. We also plan to repeat the entire experiment using participants with prior experience in analyzing mammograms (i.e. radiology trainees and residents). Finally, by combining computer vision techniques with the scanpath data of the expert that we previously recorded, we plan to explore the feasibility of an automatic scanpath predictor algorithm for mammograms.

3.7 Conclusion

Eye-tracking, mobile devices and AR have essentially created a new interactive, educational sphere, that should be exploited. All three technologies have really blossomed in recent years and are now primed, with some careful research, to create valuable, creative experiences for a wide variety of applications. Long-term, mobile AR applications are expected to have a major impact on society, infiltrating many application domains including education, tourism, engineering, defense and medicine. For example, children afflicted with Attention Deficit Disorder or mild autism could benefit from systems that help communication skills by directing gaze.

Chapter 4

High Level Saliency Prediction and its Applications in Computer Graphics

4.1 High Level Saliency Prediction for Smart Game Balancing

This section has been adapted from (Koulieris, Drettakis, Cunningham, & Mania, 2014a)

Successfully predicting visual attention can significantly improve many aspects of computer graphics: scene design, interactivity and rendering. Most previous attention models are mainly based on low-level image features, and fail to take into account high-level factors such as scene context, topology or task. Low-level saliency has previously been combined with task maps, but only for predetermined tasks. Thus the application of these methods to graphics – e.g. for selective rendering – has not achieved its full potential. In this paper we present the first automated high level saliency predictor incorporating two hypotheses from perception and cognitive science which can be adapted to different tasks. The first states that a scene is comprised of objects expected to be found in a specific context as well objects out of context which are salient (scene schemata) while the other claims that viewer's attention is captured by isolated objects (singletons). We propose a new model of attention by extending Eckstein's Differential Weighting Model. We conducted a formal eye-tracking experiment which confirmed that object saliency guides attention to specific objects in a game scene and determined appropriate parameters for a model. We present a GPU based system architecture that estimates the probabilities of objects to be attended in real-time. We embedded this tool in a game level editor to automatically adjust game level difficulty based on object saliency, offering a novel way to facilitate game design. We perform a study confirming that game level completion time depends on object topology as predicted by our system.

4.1.1 Introduction

The prediction of attention can significantly improve many aspects of computer graphics and games. For example, image synthesis can be accelerated by reducing computation on non-attended scene regions (Cater, Chalmers, & Ward, 2003); attention can also be used to improve LOD (S. Lee, Kim, & Choi, 2009). Another interesting case is computer game design. Many game genres rely on a search or target detection task to solve riddles or find game objects. If attention can be automatically predicted, several tasks in game design would be simplified. For example adjusting the difficulty of a game level could be facilitated by relocating objects estimated to attract attention (Feil & Scattergood, 2005).

Existing visual attention models such as Feature Integration Theory (FIT) are mostly driven by low level image features such as contrast, luminance and motion (A. M. Treisman & Gelade, 1980b). FIT is a commonly used model of attention in computer graphics (Itti & Koch, 2001a; Longhurst, Debattista, & Chalmers, 2006). However, it often fails to predict saccadic targets (Borji & Itti, 2013) because high-level properties such as scene semantics and task strongly affect the planning and execution of fixations (J. M. Henderson & Hollingworth, 1999; Einhäuser, Spain, & Perona, 2008; Borji & Itti, 2013). There has

been previous work accounting for task when modelling goal-oriented attention in computer graphics, but the task has always been predetermined (Cater et al., 2003; Sundstedt, Chalmers, Cater, & Debattista, 2004; Sundstedt, Debattista, Longhurst, Chalmers, & Troscianko, 2005).

Our goal is thus to develop an automated high-level saliency predictor which can be adapted to different tasks. In this paper we present the first such predictor incorporating two hypotheses from perception and cognitive science: the *scene schema hypothesis* and the *singleton hypothesis*. The scene schema hypothesis states that a scene is comprised of objects we expect to find in a specific context and salient objects that are not expected in a scene (see Figure 4.1) (Bartlett, 1932; J. M. Henderson, Weeks Jr, & Hollingworth, 1999; A. D. Hwang, Wang, & Pomplun, 2011). The singleton hypothesis states that the viewer's attention is ordinarily captured by stimuli that are locally unique in a basic visual dimension such as color, orientation, etc. i.e. isolated (Theeuwes & Godijn, 2002). In our work, the singleton state is also a context dependent measure not purely image-driven: Figure 2 shows that the spatially isolated vase attracts attention, though not salient in terms of color.

We propose a new model by incorporating the schema (Bartlett, 1932; J. M. Henderson et al., 1999; A. D. Hwang et al., 2011) and singleton (Theeuwes & Godijn, 2002) hypotheses into the Differential-Weighting Model (DWM) (M. P. Eckstein, 1998; M. P. Eckstein, Shimozaki, & Abbey, 2002; M. P. Eckstein, Drescher, & Shimozaki, 2006) that employs Bayesian Priors. To find the parameters of this model, we perform several perceptual experiments, which also verify that high-level saliency guides attention. Using this new model, we estimate the posterior probability that a viewer will fixate an object based on high-level contextual features, independent of the viewer's task (M. P. Eckstein et al., 2006). We use our automated high level saliency predictor to facilitate game level balancing, offering a novel way to ease game design.
We make three primary contributions:

- We propose a new model to account for high-level object saliency as predicted by the scene schema and singleton hypotheses by extending the Differential Weighting Model using Bayesian priors.
- In three perceptual experiments, we verify that high-level saliency guides attention and we obtain perceptual parameters to calibrate our model.
- We develop a tool based on the model to automatically predict high-level saliency in real-time. We then validate the tool's efficacy in helping to adjust game difficulty in a game-level editor.

We used a modern game engine for our experiments, our game level design and its validation. This choice underlines the relevance of our results for realistic use cases.

4.1.2 Related Work

Visual Attention Visual perception can be thought of as the active extraction and manipulation of environmental information. The visual perception pipeline starts with low-level processes which extract simple image regularities such as edges or color (Marr, 1982). Subsequently, mid-level processes combine these properties to form higher-level features such as the shape of an object (Shipley & Kellman, 2001). Finally, high-level processes map these mid-level features to meaning and semantics (Palmer, 1999). To efficiently concentrate the limited brain resources of the mid- and high-level processes on those few low-level features that are likely to be important, the human brain is equipped with a selection mechanism known as focal attention. Some low-level features such as edges can automatically attract focal attention in an almost reflex-like fashion (Koch & Ullman, 1987). Likewise, mid- and high-level features as well as goal-oriented properties can direct focal attention (J. M. Henderson et al., 1999; A. L. Yarbus, Haigh, & Rigss, 1967). For example, the contextual validity or appropriateness of an object's location will affect visual search; when looking for a chimney, usually we direct our gaze first to the rooftops. The fundamental question of how the visual system combines the influence of low-, mid-, and high-level components is a challenging research issue and remains largely unanswered (Theeuwes, 2010b). A recent review of the theories can be found in (Borji & Itti, 2013).

The most common form of focal attention model is the two-stage model, such as the popular Feature Integration Theory (FIT) (A. M. Treisman & Gelade, 1980b). In two-stage models, a privileged set of low-level features are initially extracted everywhere in an image in parallel. The focal attention mechanism then selects a few locations in the image based on these features for further processing. In the second stage, the

low level features at the selected locations are integrated and subjected to further processing in a slow, serial (i.e., one region at a time) fashion. A widely used saliency model inspired by FIT (Itti & Koch, 2001a) employs low-level features such as contrast, luminance, and motion to determine which areas are likely to attract attention. Although FIT plausibly emulates many aspects of focal attention, it has been shown that: (i) complex stimuli such as surfaces are processed simultaneously and not in a serial fashion (Nakayama, Silverman, et al., 1986), (ii) visual attention is directed to objects in a scene rather than their low level visual attributes (O'Craven, Downing, & Kanwisher, 1999) and, (iii) observers may achieve multiple simultaneous foci of attention in the visual field, not supported by FIT (Awh & Pashler, 2000). In other words, attention models based on low-level features often fail to predict saccadic targets (Borji & Itti, 2013), in part because they do not take into account high level factors such as scene context, task, or object topology (Einhäuser, Spain, & Perona, 2008; J. M. Henderson & Hollingworth, 1999; R. A. Rensink, 2000).

Two phenomena within the perception literature point to specific roles that high-level information can play in focal attention. The first – the *scene schema* effect – is based on the observation that a high proportion of objects in a scene can usually be expected to be found there. They are "consistent" with the scene. Sometimes, however, objects are in a scene or a location that is very atypical. Such "inconsistent" objects are potentially salient (see, e.g., Figure 4.1) (Bartlett, 1932). Research has shown that previously-acquired knowledge of stereotypical object placement in a scene combined with the on-going visual experience of a scene can attract focal attention (Brewer & Treyens, 1981; Bar, Ullman, et al., 1996; J. M. Henderson et al., 1999). The ratio and location of consistent and inconsistent objects in a specific context can also influence whether the scene is perceived to be congruent overall (Einhäuser, Spain, & Perona, 2008; Rayner, 2009; A. D. Hwang et al., 2011). The second effect – the *singleton effect* – refers to the finding that stimuli that are locally unique in terms of color or orientation capture attention (Figure 4.2) (Theeuwes & Godijn, 2002). Object perception is based on context-dependent processing of low-level variables i.e. pixels, therefore the singleton state is a high level semantic property of spatially isolated objects.

More recently, a number of single stage models have been proposed, which are very effective at describing visual attention. For example, Eckstein has proposed a single-stage model of attention called the Differential-Weighting Model (DWM), (M. P. Eckstein, 1998; M. P. Eckstein et al., 2002, 2006) which incorporates both low-level features as well as prior knowledge about scene context. The DWM models attentional processing using physiological noise in brain neurons and Gaussian combination rules. Contextual information in the DWM is embodied in the Bayesian priors provided to the model beforehand. For example, when searching for a chimney in a picture that contains a house, the visual elements depicting the roof of the house are given a higher prior probability than other scene elements. DWM has never been used to predict high-level saliency or gaze patterns in interactive Virtual Environments (VEs) incorporating scene schemas and singletons.

Attention in Computer Graphics In an effort to predict attention in pre-determined task areas, it has been shown that task importance maps may be used to accelerate rendering by reducing quality in regions that are unrelated to a given task (Cater et al., 2003). Selective rendering guided by a FIT-based saliency model renders perceptually important parts of a scene in high quality while the remaining areas of the image are rendered at lower quality, thus saving in computational cost (Longhurst et al., 2006). As mentioned, FIT only uses low-level image characteristics. Other research has combined task maps with a low-level saliency map and validated the results using eye-tracking (Sundstedt et al., 2004, 2005). Predicting gaze behavior in games may be used to optimize the distribution of computing resources (Sundstedt, Stavrakis, Wimmer, & Reinhard, 2008). Saliency models and task related data have been linearly combined to track visually attended objects in a VE in task-specific areas (S. Lee et al., 2009). Task relevant gaze behavior associated to first-person navigation in a virtual environment has been estimated by combining bottom-up and top-down components to compute user gaze point position on screen (Hillaire et al., 2010). Attention in games may also get manipulated. A guiding principle and method based on the Guided Search theory (J. M. Wolfe, 1994) has been proposed to direct attention to target items that should be noticed by an observer in a video game e.g. an advertisement. When a frequently searched game object is modified to share perceptual features such as color or orientation with a target item, the item will attract attention (Bernhard, Zhang, & Wimmer, 2011). Saliency models have been employed to animate the gaze behavior of virtual characters (Oyekoya, Steptoe, & Steed, 2009) and crowds (Grillon & Thalmann, 2009).

Although task-based saliency estimations competently predict salient regions in pre-determined task-specific areas (Cater et al., 2003), the challenge is to estimate salient regions in all areas of a scene for different

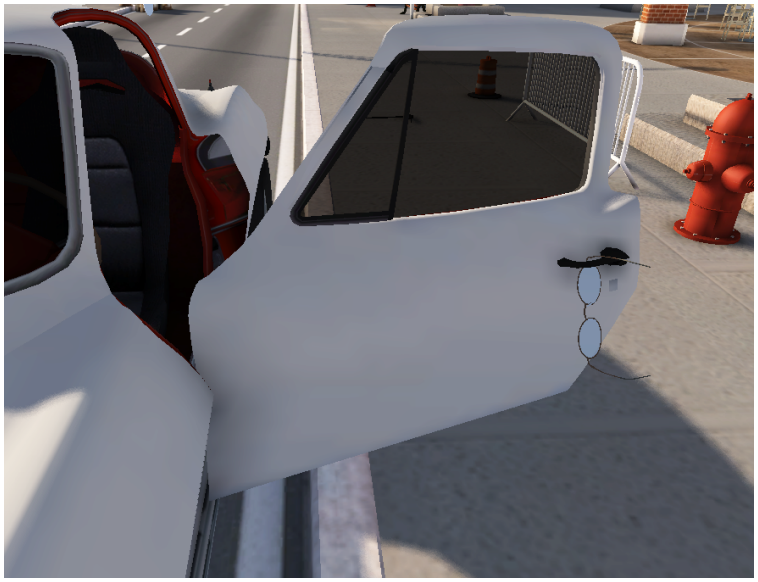


Figure 4.1: The spectacles attract attention as they are *inconsistent* with the car door context.

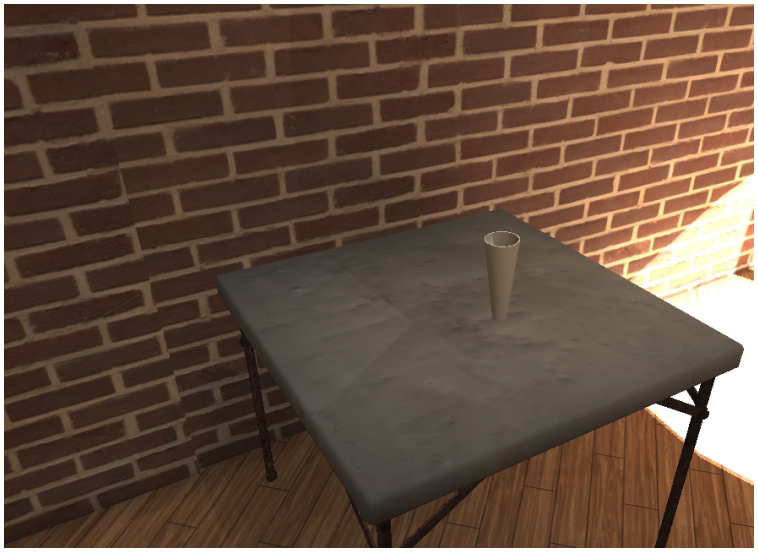


Figure 4.2: The spatially isolated vase attracts attention as it is a *singleton* object.

tasks via an integrated model. Research in interactive VEs has confirmed that attention is influenced by the semantic context of objects in the form of scene schemas (Mania, Robinson, & Brandt, 2005; Mourkoussis et al., 2010; Zotos et al., 2009). In one step towards implicitly modelling high-level effects, machine learning techniques have been applied to eye tracking data in order to train a model to detect salient regions in a pre-defined set of static photographs (Judd, Ehinger, Durand, & Torralba, 2009b). A pipeline to derive gaze prediction heuristics from eye-tracking data for 3D Action Games has been proposed (Bernhard, Stavrakis, & Wimmer, 2010). To date, a model that explicitly links in a physiologically plausible manner experimental outcomes on attention with object saliency is missing.

We propose an innovative visual attention model based on DWM which takes into account high level information about the context of the scene. Our new model can be directly used in computer graphics. We validate our theoretical hypotheses in a formal perceptual study involving game level balancing.

Attention in Computer Games Gameplay greatly depends on attention deployment (Sundstedt, Bernhard, Stavrakis, Reinhard, & Wimmer, 2013). Eye tracking data has revealed that players playing first person shooter games tend to concentrate on the center of the screen searching for enemies while in an Action-Adventure game players mostly explore the entire screen for game props to advance the gameplay (El-Nasr & Yan, 2006).

Player enjoyment is crucial for the success of a computer game. An enjoyable/optimal experience, also termed *flow*, is shown to be so satisfying that players take pleasure in the game with little concern for what they will get out of it (Czikszentmihalyi, 1990). Enjoyable experiences in games arise primarily from challenge (Sweetser & Wyeth, 2005). Challenge refers to the ability of a game to be sufficiently intriguing and match the player's skill level (Pagulayan, Keeker, Wixon, Romero, & Fuller, 2003; Desurvire, Caplan, & Toth, 2004). Improper challenge levels provoke anxiety in a discouragingly hard game or apathy in a boringly easy game (Johnson & Wiles, 2003).

Looking for an object is a common task in Adventure or Action-Adventure video games, often guiding level advances. The time spent searching for an object in a game should be in proportion to the advantage it conveys in game play. Designers mostly rely on their experience and instinct while calculating cost/benefit ratios by manually placing objects and obstacles in their levels (Pagulayan et al., 2003). Multiple rounds of Play-Testing and observation can stabilize choices in a level (Sweetser & Wyeth, 2005). However, because players' abilities vary and play-testers are not abundant to every game designer, a sophisticated approach such as the model we propose, that guides automatic object manipulation and game balancing based on high-level visual attention is crucial.

4.1.3 Modeling High Level Saliency

In this section, we present our new model of high-level attention. Before presenting our new model, we first describe the DWM. We then explain how we extended DWM by encoding the interaction of schemas and singletons based on the Bayesian priors of the original model.

4.1.4 The Differential-Weighting Model

The Differential-Weighting Model (DWM) (M. P. Eckstein, 1998; M. P. Eckstein et al., 2002, 2006) estimates the interaction between visual evidence concerning a target in a scene and Bayesian prior probabilities indicating expectation and context of a scene. By combining sensory data with existing knowledge it calculates the posterior probability that a location will be fixated in a visual search task and thus predicts saccadic targeting.

DWM assumes that when searching for a target, each location in a scene elicits neuronal activity in relevant sensory units of each visual feature. This response is subject to Gaussian independent neutral noise, i.e. the outcome of the perceptual processing of this response is probabilistic. When a sensory unit is tuned to observe a specific feature, it responds at a higher rate when the observed feature is present. Neurons are subject to internal noise and have a response following a Gaussian distribution (Tolhurst, Movshon, & Dean, 1983). After many trials, Figure 4.3 depicts the internal response probability density functions for noise-alone (left curve) and for signal-plus-noise trials (right curve). The model calculates the ratio of the joint likelihood of observing the feature's neural responses in each image region given that the target is

present and the joint likelihood of observing the feature's responses given that the target is absent according to a selected probability. This noisy response is then weighted by context effects encoded in Bayesian priors relevant to specific stimuli. The Bayesian priors embody the probability of these stimuli to co-occur with other highly visible visual features of the image.

For each image frame f and each visual field location (x, y), each sensory unit responds in a noisy manner for each feature λ_j. DWM calculates the likelihood $l_{j,x,y,f}$ of observing the response λ_j given the presence of the target's j^{th} feature at that location and the likelihood of the response given the absence of the feature. The response has a Gaussian distribution (Tolhurst et al., 1983) with a mean of d'_j and a standard deviation σ. The likelihood $l_{j,x,y,f}$ that the j^{th} sensory unit takes a value $\lambda_{j,x,y,f}$ given the presence of the target's j^{th} feature at (x,y) on frame f is then

$$l_{j,x,y,f}(\lambda_{j,x,y,f}|s) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \left(\frac{(\lambda_{j,x,y,f} - d'_j)^2}{2\sigma^2} \right) \tag{4.1}$$

s stands for signal and denotes the presence of the target.
The likelihood that the j^{th} sensory unit takes a value λ_j given the absence of the target's j^{th} feature is

$$l_{j,x,y,f}(\lambda_{j,x,y,f}|n) = \frac{1}{\sqrt{2\pi\sigma^2}} exp - \left(\frac{(\lambda_{j,x,y,f})^2}{2\sigma^2} \right) \tag{4.2}$$

n stands for noise and denotes the absence of the target.
A likelihood ratio LR (Green, Swets, et al., 1966) can be calculated as

$$LR_{j,x,y,f} = \frac{l_{j,x,y,f}(\lambda_{j,x,y,f}|s)}{l_{j,x,y,f}(\lambda_{j,x,y,f}|n)} = exp \left(\frac{\lambda_{j,x,y,f} d'_j - 0.5 d'^2_j}{\sigma^2} \right) \tag{4.3}$$

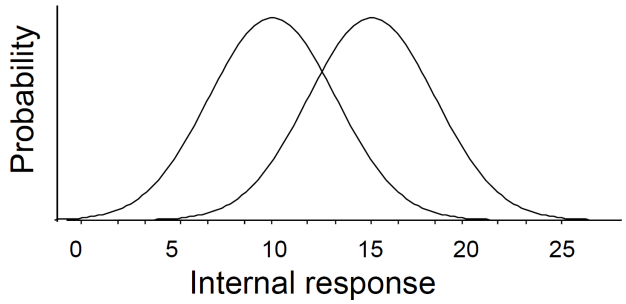


Figure 4.3: Internal response probability density functions for noise-alone (left curve) and for signal-plus-noise trials (right curve).

4.1.5 A New High-level Attention Model

We propose a new model by integrating high-level information implied from semantic (schema inconsistency, Figure 4.1) or physical (singletoness, Figure 4.2) context represented by Bayesian priors in the DWM. The internal response associated with high level saliency components is also subject to noise. We assume that dedicated, high-level sensory units are analogous to low-level sensory units. The high-level units fire at their highest rate when fed with the correct high-level feature, much as a low-level edge-detection unit reacts highest when an edge with the proper orientation is presented (M. P. Eckstein, 1998; M. P. Eckstein et al., 2002, 2006). Whether the neural mechanism underlying a high-level sensory unit is a single neuron or a cluster of neurons does not matter. What matters is that there is an internal (neural) state reflecting whether this high-level feature is present or not. For example, we propose a sensory unit that monitors the degree to which an object is isolated. Such a unit would fire when a singleton object is in the field of

view (Steinmetz et al., 2000).

We extended the original DWM equations to describe two high-level sensory units tuned to schema inconsistencies and the singleton state of objects. Equations (1) – (3) assume that the internal response generated by the presence of each visual feature is known a priori. Since neuronal response strength is unknown concerning scene schemata and singletons, we alter the DWM and instead calculate the posterior probability that the target is present at each pixel as a sum of K different feature strengths d'_k associated with scene schemata and singletoness.

$$P_{semantic,x,y,f}(s|\lambda) = \sum_{k=1}^{K} LR_{semantic,x,y,f,k} \tag{4.4}$$

$$P_{physical,x,y,f}(s|\lambda) = \sum_{k=1}^{K} LR_{physical,x,y,f,k} \tag{4.5}$$

We then average the components (4), (5) using weights $w_{semantic}$ and $w_{physical}$ that we obtain from perceptual studies (see Section 4.1.6) to calculate the posterior probability that a location will be attended. A linear combination of components is a common practice in saliency detection algorithms (Frintrop, Rome, & Christensen, 2010b).

$$P_{x,y,f} = w_{semantic} P_{semantic,x,y,f} + w_{physical} P_{physical,x,y,f} \tag{4.6}$$

As an example consider a bar counter. A coffee mug which is consistent with the context and a medical kit which is inconsistent with the context are shown in Figure 4.4 (top). Consider a sensory unit that tracks schema inconsistencies. The $\lambda_{semantic}$ of the image regions corresponding to the medical kit is higher than the λs associated to the mug and counter. The $\lambda_{semantic}$ communicates a subjective rating of consistency, e.g. the higher the number, the more inconsistent the object is in relation to the context (Figure 4.4 bottom left). Let us assume that $K = 1$, $d'_{semantic} = 0.6$ and $\sigma = 0.2$. Because the medical kit is inconsistent, we assume that $\lambda_{semantic} = 1.0$, similarly for the mug $\lambda_{semantic} = 0.16$, for the bar counter $\lambda_{semantic} = 0.22$ because they are both highly consistent. The likelihoods ratios of observing the medical kit, mug and counter are $LR_{medikit} = 36315.5$, $LR_{mug} = 0.1$, $LR_{counter} = 0.3$ respectively as derived from equation (3). The schema inconsistency unit would then estimate the medical kit as the most salient (Figure 4.4 bottom right). Similarly, $\lambda_{physical}$ is used to calculate the likelihood ratios of observation based on whether an object is placed as singleton.

Figure 4.4: A bar counter context (top), $\lambda_{semantic}$ visualization (bottom left) and highest $LR_{semantic}$ highlighted (bottom right).

4.1.6 Real-time evaluation of High Level Saliency Components

We examined the real-time effect of singleton and scene schemata on gameplay for two reasons:

- The role of scene schemata and singletons in interactive, synthetic environments is unknown, even though their effects are well-documented for target detection displays or static real photographs (J. M. Henderson et al., 1999; R. A. Rensink, 2000; Einhäuser, Spain, & Perona, 2008; Theeuwes & Godijn, 2002).
- Our extension of the DWM requires an empirical classification of objects in relation to scene schemata and the determination of the weighting factors w_j that signify the interaction between semantic (scene schemata) and physical context (singletons).

Inspired by Adventure games (Ju & Wagner, 1997), a suitable game genre to apply our method, we designed an environment that allows us to investigate the impact of high-level saliency on visual attention & gameplay and recorded the time it took to search for plot-critical objects. The storyboard was based on the popular video game L.A. Noire™, a 2011 Action-Adventure neo-noir crime video game developed by Team Bondi™ and published by Rockstar Games™. A scene depicting a Coffee Shop inspired by the "Driver's Seat" case of the game was heavily modified to include multiple areas representing a car schema and a cafeteria schema inclusive of sub-schemata representing a coffee shop counter and a lounge loft. We systematically controlled the semantic and physical states of plot-critical objects. Each object could be in a schema-consistent or a schema-inconsistent location, and could be in either a singleton state (positioned by itself) or a compound state (positioned in cluttered surroundings) (please see accompanying video).

Experiment 1: Defining object consistency

Here, we empirically classify scene objects as either consistent or inconsistent in relation to the context of each part of the scene. Specifically, a list of 50 objects was assembled and given to 21 graduate students (14 male, 7 female). Each participant used a 7-point Likert scale to rate how likely each item was to appear in a given scene. A rating of 7 meant that the object was very much expected to be in that location and 1 meaning the object was very much not expected. Half of the objects were tested in the Coffee Shop counter context and the other half were tested in the car context. We then selected a set of consistent objects from the high end of the scale and a set of inconsistent ones from the low end (based on the approach used in (Brewer & Treyens, 1981)). The classification of objects in relation to scene schemata is independent from a specific game scenario, i.e. a teapot is consistent with a kitchen context irrespectively of a background story. A taxonomy of common objects in relation to scene schemata that can be used in any game will be provided as part of the production level version of our system.

Experiments 2 & 3: Determining the Roles of Semantic and Physical Context

In Experiments 2 and 3 we examine the effect of physical (singletoness) and semantic (consistency effects) manipulations on game task completion time for two common tasks appearing in (Action-)Adventure games. In both tasks, the same general scenario was used: "*Adrian Black, a married man and a barista at the Coffee Shop decides to start a new life with his customer Nicole staging his own murder to cover a getaway with her.*" Participants were instructed in both tasks to find three decisive objects as quickly and as accurately as possible in order to solve the mystery (Figure 4.5). Experiment 2 used a *Search* task (participants knew exactly what they were searching for). Experiment 3 used a *Non-Search* task (participants did not know what they were searching for, and as such were exploring the environment with less of the specific purpose). Our two main predictions are:

- Singleton objects will require less time to be recovered compared to objects in compound state because they capture attention no matter what the task is (Theeuwes & Godijn, 2002).
- When searching for an object, consistent locations will attract attention and therefore will require less time to be recovered than inconsistent locations. When not searching, on the other hand, objects at inconsistent locations should attract attention and therefore will require less time to be recovered compared to consistent locations (M. P. Eckstein et al., 2006).

Figure 4.5: One of the desicive objects, the spectacles, as positioned in different conditions: Consistent/Compound (left) vs Inconsistent/Singleton (right).

Method

Each of the two main factors (Semantic Context and Physical Context) had two levels, which were factorially combined to produce four experimental conditions: Consistent/Compound, Inconsistent/Compound, Consistent/Singleton, and Inconsistent/Singleton object placement. The objects were positioned so as to maintain constant navigation time while reaching them across conditions and on similar visual angles within the VE. The four conditions above were the same for both experiments. A between-participants design was used, meaning that each person participated in only one experiment and in only one experimental condition.
Participants A total of 80 participants (56 male, 24 female; ages between 21 - 33) were recruited from the undergraduate and research population of our institution and were rewarded with pastry for their participation. All participants were familiar with first person perspective navigation and had normal to corrected vision. Upon arrival, each participant was randomly assigned to one of eight groups so that each group had 10 participants. Each group participated in only one of the experimental conditions.
Procedure and Apparatus Upon arrival, the participants signed a consent form and were then allowed to practice navigating in a training scene. The participants were then informed of the experimental scenario and positioned about 60cm from a 20" flat screen monitor (screen width of 44cm) at a resolution of 1680x1050. The game environment was rendered in real-time at a 60Hz constant refresh rate. First person viewing mode was used for navigation. The virtual camera was positioned at the level of the eyes of the subject's avatar which was 1.80m in height. The avatar had three degrees of displacement freedom. Yaw and pitch angles of the camera were controlled with the mouse, while walking was controlled with the arrow keys of the keyboard. Task completion time as well as inspection start/end timings indicated by a mouse over a possible clue, collect attempts, collected (decisive or not) objects were stored in a database along participants' age and gender.

Results

We subjected the completion times to a Multiple Linear Regression (MLR) analysis which, like the ANOVA, is a subclass of *general linear modelling*. Unlike the ANOVA, a linear regression also provides an explicit, quantitative model of how the different experimental factors affect performance along with the relative importance of the different factors (Cunningham & Wallraven, 2011). This information is critical for deriving the DWM weights.

In MLR, the line $y = m_1x_1 + \cdots + m_nx_n + b$ is fit to the data, with y being the participants' performance (e.g., task completion time) and each x_i being an experimental factor (e.g., physical or semantic context) and b being the intercept. Since our two factors are categorical, they must be *dummy coded*. We gave Compound a value of 0 and Singleton a value of 1. Likewise, Inconsistent and Consistent were set to 0

Coefficients	Estimate Time	p-value
Intercept	158.962	< 0.0001
+Singleton placement	-81.309	< 0.0001
+Consistent placement	-18.381	0.003
+Joint Term	22.190	0.055

Table 4.1: The regressions coefficients for each factor and whether that factor significantly contributed to the overall model are listed for a Search task.

Coefficients	Estimate Time	p-value
Intercept	153.008	< 0.0001
+Singleton placement	-67.111	< 0.0001
+Consistent placement	11.944	0.039
+Joint Term	-32.407	0.025

Table 4.2: The regressions coefficients for each factor and whether that factor significantly contributed to the overall model are listed for a Non-Search task.

and 1, respectively. Each regression coefficient m_i indicates how many seconds faster a unit change (i.e., from 0 to 1) in the factor x_i will cause the completion time to be. Critically, the ratio of the mean squared prediction error of a model to the variance in completion time is directly related to the pearson correlation coefficient (Cunningham & Wallraven, 2011) and indicates how much of the variance in completion time can be "explained" or predicted by the change in the independent variables. We will use this relative predictive values to derive the DWM weights.

Experiment 2: Search Task On average, participants needed 64.81, 72.10, 135.03 and 164.5 seconds to complete the Singleton/Consistent, Singleton/Inconsistent, Compound/Consistent, and Compound/Inconsistent conditions, respectively (see Figure 4.6). Regressing physical context onto completion time yields a model that explains 80.7% of the variation in completion time. This is a significant amount, $F_{1,38} = 159.1, p < .001$, showing the significant effect of physical context. There was also a significant effect of semantic context: a two predictor model regressing both physical and semantic context onto completion time explains 84.8% of the variance. This increase in predictive power of 4.1% is statistically significant, $F_{1,37} = 10.068, p < 0.0031$. Finally, the interaction between physical and semantic context was marginally significant: adding a term to capture the variance jointly explained by semantic and physical context – while controlling for multicolinearity – explains an additional 1.5%, $F_{1,37} = 3.9621, p < 0.055$. The intercept, regression coefficients and statistical significance of each predictor in the two and three predictor models can be seen in Table 4.1. As can be seen in the table, the two predictor model predicts that performance in the Compound/Inconsistent condition should be 158.962 (the intercept) which is close to the actual value of 164.5. Changing from compound to singleton should speed up performance by 81.309 seconds (the regression coefficient for physical context), and changing from inconsistent to consistent should speed up performance by 18.381 seconds. Thus, performance in the Singleton/Consistent condition is predicted to be 59.272, which matches the actual value of 64.81 well.

Experiment 3: Non-Search Task On average, participants needed 89.74, 94, 173.05 and 144.90 seconds to complete the Singleton Consistent, Singleton Inconsistent, Compound Consistent, and Compound Inconsistent conditions, respectively (see Figure 4.7). The effect of physical context was again significant; a single predictor model explains 77.7% of the variance, a statistically significant amount, $F_{1,38} = 132.1, p < .001$. Semantic context was also significant; the two predictor model explained 80.2% of the variance, a statistically significant increase of 2.5%, $F_{1,37} = 4.578, p < 0.04$. The interaction was also significant; the three predictor model explains 84.7% of the variance, an increase of 4.5%, $F_{1,37} = 3.258, p < 0.003$. The intercepts and significance of the three predictors can be seen in Table 4.2.

4.1.7 Discussion

Both semantic and physical context play a statistically significant role in attention deployment, with physical context playing the dominant role. Moreover, an object is often inconsistent with its surroundings (and thus

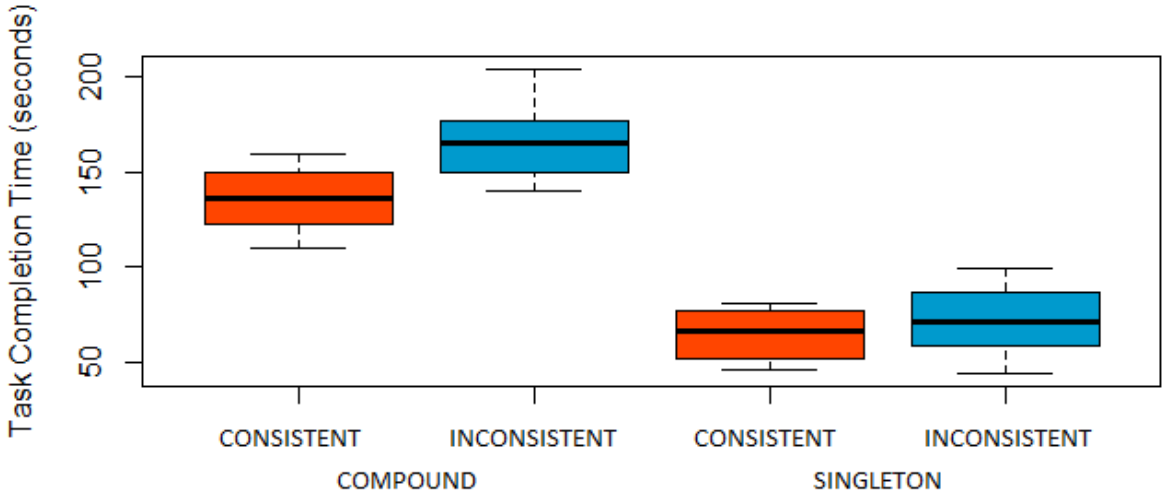


Figure 4.6: Task completion time distribution in a Search task. The thick, horizontal line in each box represents the median for that condition. The colored box around the median represents the middle quartiles and the outer bars represent the extremes.

will probably grab attention) but neither in a singleton state nor salient in terms of low level features. In such cases, the scene schemata theory can predict its prominence. In agreement with our first prediction, placing an object in a singleton state decreased task completion time. The two predictor model indicates that performance in the singleton conditions is about 49% of that in the compound conditions for Search tasks, and about 59% for non-search tasks. In agreement with the first part of our second prediction, consistency decreases task completion time for a Search task. The significant interaction for Non-Search tasks, however, means that the effects of semantic was dependent upon physical consistency: inconsistent locations were only faster for compound objects. Contrary to prediction, inconsistency increased search time in a Non-Search task for singleton objects.

4.1.8 Model Initialization

We used the results of the previous two experiments to derive weighting factors w_j for each dimension. In a search task, a two predictor model explained 84.8% of the variance, with object singletoness explaining 80.7% and schema consistency 4.1%. Thus, $w_{physicalSEARCH} = 0.95$ (80.7% out of 84.8%) and $w_{semanticSEARCH} = 0.05$. In a Non-Search task, object singletoness explained 77.7% of the total 80.2%, giving us $w_{physicalSEARCH} = 0.97$ and $w_{semanticSEARCH} = 0.03$.

In order to calculate the likelihood values associated to the scene schema hypothesis we compare the associated scene schema of each examined object determined in Experiment 1 against the scene schemata associated with the objects that surround it. We define an object neighbourhoud of radius N as a multiple of the examined object's radius. We define c the count of objects residing in this neighbourhoud and m the count of items tagged with the same schema inside the neighbourhoud.
We then define $\lambda_{semantic}$ as:

$$\lambda_{semantic} = \frac{c - m}{c} \tag{4.7}$$

Inconsistent objects signified by their varied schema relatively to their surroundings have greater $\lambda_{semantic}$ values than consistent objects.

In order to calculate the likelihood values associated to the singleton hypothesis, we both examine the number of neighbours for each examined object and employ the available image depth information. In particular, we can use the spatial derivatives to estimate the magnitude of the depth gradient. This operator indicates how distinct an object is from its environment and is a strong indication of whether it is a singleton.

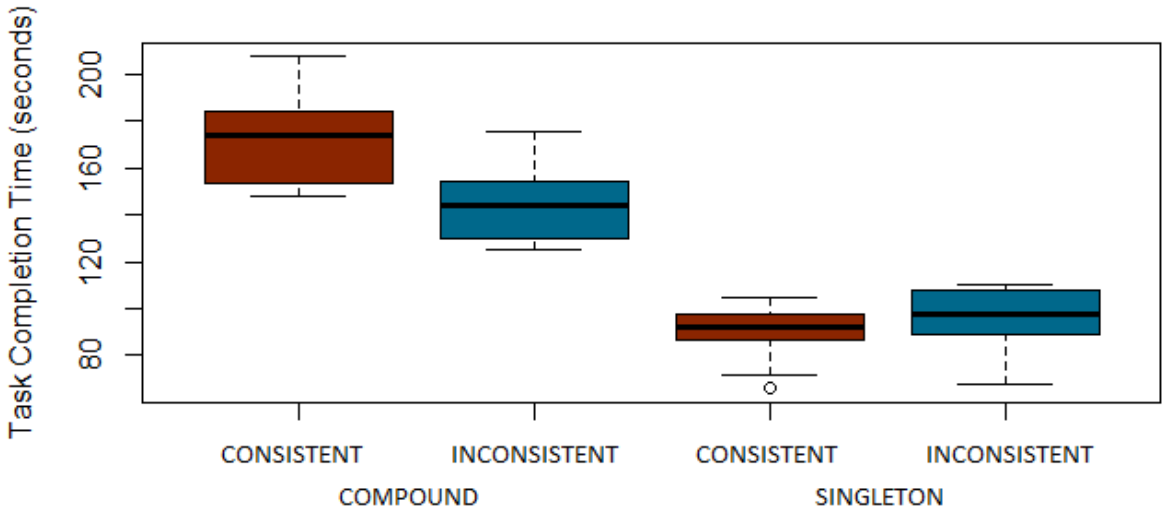


Figure 4.7: Task completion time for a Non-Search task.

We thus define $\lambda_{physical}$ as:

$$\lambda_{physical} = \frac{1}{1-c} \times \sqrt{(\frac{\partial f}{\partial x})^2 + (\frac{\partial f}{\partial y})^2} \tag{4.8}$$

4.1.9 Implementation and Game Balancing

In this section we describe a GPU implementation of our model and its integration in a game engine to assess game level difficulty. The efficiency of our model in predicting attention deployment is evaluated in Experiments 4 & 5.

4.1.10 GPU based implementation

We developed a plug-in for Unity 3D^{TM} game engine which we call High Level Saliency Modeler (HLSM). HLSM highlights objects expected to attract attention by estimating in real time the posterior probability term (Equation 4.12) of our new high level attention model in a shader (Figure 4.1.11) (please see accompanying video). Equations (1)-(5) are supplied with semantic consistency and object singletoness information in terms of the $\lambda_{physical}$ and $\lambda_{semantic}$ variables as determined by the experiments. The $\lambda_{semantic}$ (Equation 7) and $\lambda_{physical}$ (Equation 8) are calculated at runtime by both quering the scene graph and utilizing an edge detection kernel run over the depth buffer. The obtained likelihood ratio sums are then combined according to the w_j factors obtained from the regression analysis applied to the experimental task completion timings (see Section 4.1.8). The d', σ and K values are user controlled via the system's user interface. Manipulating these parameters either increases or decreases the system's sensitivity to saliency resulting in more or fewer objects to be highlighted as salient respectively (Figure 4.1.11).

The shader approach offers view-dependent estimations i.e. an object may or may not appear as singleton depending on the viewpoint. Additionally, the linearity of the likelihoods calculated allows for linear quantitative measurements. For instance, "an object x is more inconsistent than object z by a factor of q". This offers rich information about the semantic context of objects as opposed to the previously defined binary definition of an object being characterized as either consistent or inconsistent (Zotos et al., 2009).

4.1.11 Game Level Editing

Game balancing is a meaningful application of high level saliency modeling. Plot-critical objects are placed in their respective locations by game designers to achieve a purpose: Ease or make it difficult for the player when searching for them depending on the plot. Placing objects far from expected locations is standard

in game balancing (Feil & Scattergood, 2005). Integration of a high level saliency model in a game level editor can assist the level artists by highlighting salient objects. Designers using the proposed editor are able to reposition or tint props to make them less/more visible in real-time. This way, designers modulate the search-cost/benefit curve for easier or harder object recovery in Adventure or Action-Adventure games. When working with our tool, the game level designer proceeds as normal to place game objects as desired. The designer observes saliency visualization and examines the attention prediction for the current view. The current view or object placement may then be modified and high level saliency can be re-assessed in real-time. The overhead of investigating the attention predictions is minimal since the game level designer may save on time by not needing to elaborate on suitable locations for prop placement depending on the current game difficulty level that is developed. We used the GPU implementation of our model to guide object placement in a game level editor in order to adjust game difficulty. Our plug-in works in parallel with the editor, allowing the game designer to play-test the level while designing it.

Experiments 4 & 5: Evaluation of the implementation

We designed an experiment to evaluate the efficiency of our model in predicting attention deployment by examining its effect on task completion time and by acquiring eye-tracking data. Since our tool is intended to be used by game level designers when creating game levels, the evaluation also indicates the model's potential as a means to adjust game level difficulty.

Design

We created four game levels corresponding to two experimental conditions (*Easy*/*Hard*) of a Search and a Non-Search Task. The placement of three critical objects was manipulated to systematically alter game difficulty. Our model implementation (HLSM) assisted object placement by highlighting objects that were expected to pop out in a Search task for the first two conditions (*Experiment 4*) and in a Non-Search task for the last two conditions (*Experiment 5*). Figure 4.1.11 shows a vase at a consistent/singleton layout expected to attract attention in a Search Task and thus marked as red by HLSM. When the vase is placed on the chair therefore being at an inconsistent/compound location it is not expected to pop-up in a Search Task. When objects pop out we expect a shorter task completion time, thus the easy level for the Search task was created by placing consistent objects at a singleton state in the scene. A hard game level expected to be completed slower was created by placing inconsistent objects at a compound state (Table 4.1). In relation to the Non-Search task the easy level was created by placing consistent objects at a singleton state and the hard level by placing consistent objects at a compound state expected to have the fastest/slowest recovery times respectively (Table 4.2). In all cases we use our saliency modeller, which indicates the appropriate configurations. We used the Saliency Toolbox (Walther & Koch, 2006) to ensure that the requested objects exhibited a minimum low level saliency (Figure 4.1.11). Constant navigation time to the individual objects was maintained regardless of location. Similar visual angles within the VE were maintained for all objects.

Participants and Apparatus

Forty participants (34 male, 6 female; mean age 23) were split in four groups; 10 played the easy Search task level, 10 played the hard Search task level, 10 played the easy Non-Search task level and the rest played the hard Non-Search task level. For the Search task participants were instructed to find three specific objects. For the Non-Search task participants observed the VE to identify three unknown objects that were indirectly described, for example, "identify objects necessary for a car trip" (Figure 4.1.11). For both tasks participants were instructed to find the objects as quickly and as accurately as they could. Each subject participated in only one of the experimental conditions. The VEs were presented in stereo at SXGA resolution on an NVIS nVisor SX111 Head Mounted Display with a Field-of-View of 102 degrees horizontal. An InterSense InertiaCube3, three degrees of freedom head tracker was utilized for rotation and a gamepad for translation. Attached to the HMD was an eye-tracker by Arrington Research recontructing the subjects eye position through the Pupil-Center and Corneal-Reflection method at a rate of 30Hz. The eye tracking was performed to the dominant eye of each subject.

Results

Experiment 4: Search Task An independent-samples t-test was conducted, revealing a significant difference between easy (M=42.83, SD=11.83) and hard (M=82.2, SD=21.88) level completion times, $t(9) = -4.54$, $p < 0.0001$. The easy task completion time was reduced to 52.1% of the hard task; 42.83 vs 82.2 seconds, that is consistent with the results of the regression analyses of Experiment 2: A consistent/singleton object placement is predicted to be reduced to 37% of an inconsistent/compound object placement completion time derived from 59.272 (intercept+singleton+consistency terms) vs 158.962 seconds (Table 4.1).
Experiment 5: Non-Search Task An independent-samples t-test was conducted, revealing a significant difference between easy (M=61.86, SD=17.57) and hard (M=138.35, SD=16.1) level completion times, $t(9) = -14.48$, $p < 0.0001$. The easy task completion time was reduced to 44.7% of the hard task; 61.86 vs 138.35 seconds, that is consistent with the results of the regression analyses of Experiment 3: A consistent/singleton object placement is predicted to be reduced to 39.67% of a consistent/compound object placement completion time derived from 65.434 (intercept+singleton+consistent+joint terms) vs 164.952 seconds (Table 4.2).

Discussion The reduction of task completion time in the easy conditions when compared to the hard conditions for both the Search and Non-Search tasks validate our hypothesis that game level completion time depends on object topology as predicted by our system.

Analysis of the Eye-Tracking Data

For every object in quest a Region-Of-Interest (ROI) was defined. Each ROI held metadata indicating a consistent/inconsistent placement and a singleton/compound placement of the object in relation to its surroundings. In total 9837 fixations to the ROIs were recorded. As a fixation we considered every spatially stable gaze lasting for at least 300 milliseconds (Salvucci & Goldberg, 2000).

For **Experiment 4** an independent-samples t-test was conducted on total object fixations per condition, revealing a significant difference between consistent/singleton (M=265.3, SD=15.41) and inconsistent/compound (M=182.6, SD=25.16) object placement, $t(9) = 7.45$, $p < 0.0001$.
For **Experiment 5** an independent-samples t-test was conducted on total object fixations per condition, revealing a significant difference between consistent/singleton (M=364.5, SD=44.92) and consistent/ compound (M=171.3, SD=19.04) object placement, $t(9) = 15.6$, $p < 0.0001$.

Discussion The results indicate a clear influence of context consistency in attention deployment for the Search Task. Singleton objects attracted attention in both conditions since the total number of fixations for ROIs defined for objects in a singleton state were higher for both the Search and Non-Search tasks. We aggregated fixations collected over raw eye data from all participants and visual angles in multiple heatmaps (Figure 4.1.11). Observing the heatmaps indicated that in a Search task eye gaze is directed significantly more often to consistent locations in relation to the requested object (Figure 4.1.11). In a Non-Search task the eye scan pattern spans over the entire scene, which is consistent with previous literature stating that in an Action-Adventure game players mostly explore the entire screen for game props to advance the gameplay (El-Nasr & Yan, 2006) (Figure 4.1.11). Our model implementation succesfully predicts the saliency of objects (Figure 4.1.11) that were identified as non-salient in terms of low level features (Figure 4.1.11) further validated by the eye-tracking study (Figure 4.1.11). Adjusting game level difficulty by manipulating object topology is thus feasible in Adventure or Action-Adventure games.

Conclusion & Future Work

This work presents a first attempt to devise a high level saliency predictor based on the topological relationships of objects with their surroundings and object-scene schema conformance for common tasks in (Action-)Adventure games. The framework automatically estimates attention deployment by identifying salient regions in the viewpoint. We conducted three experiments to verify that high level saliency of objects affects the time needed to find them in a VE and also obtained all the necessary weighting factors

Figure 4.8:
In a Search task, our tool highlights the vase at a consistent/singleton location signifying an easier recovery than at an inconsistent/compound location (on chair). The green hue indicates non-salient areas.

Figure 4.9:
The system's sensitivity to saliency can be adjusted, resulting in more (left) or fewer (right) objects to be highlighted as salient.

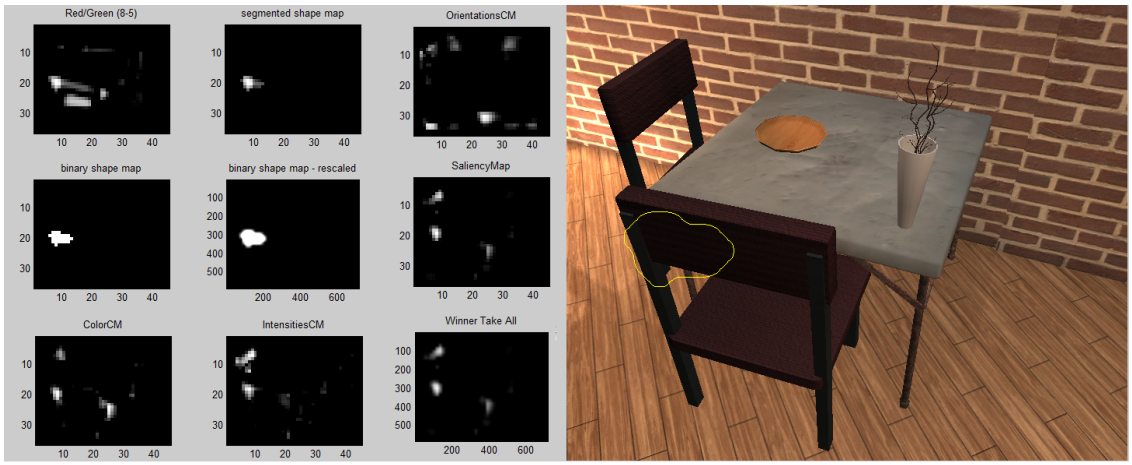


Figure 4.10:
The Saliency Toolbox (Walther & Koch, 2006) indicates that the most salient area of the image is the dark area behind the chair.

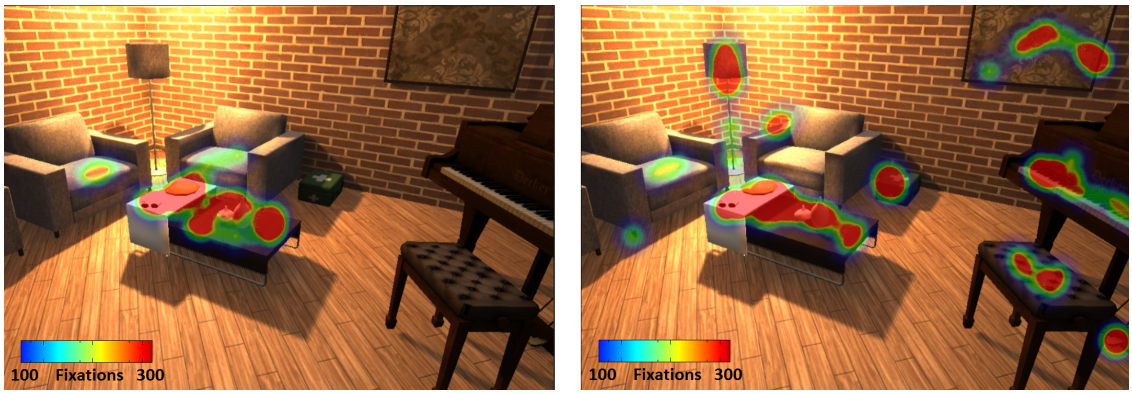


Figure 4.11:
The left image indicates fixations for a Search task where subjects were requested to find a pair of spectacles. The right image indicates fixations for a Non-Search task where subjects were requested to identify objects necessary for a car trip. Areas receiving less than 100 fixations are excluded to eliminate noise.

for our model. Finally, we developed a GPU based computational model that implements our new model incorporating high level saliency components. The system estimates the probabilities of individual objects to be foveated in real time and can be used in an innovative game level editor automatically suggesting game objects' positioning in order to adjust the difficulty of the game. The system can be adapted to additional tasks, different than the ones presented here by acquiring the necessary parameters using the methodology we presented.

We plan to include additional High Level Saliency Components in our model such as Perception of Sets. i.e. attention regarding large aggregations of similar objects as a single unit (Ariely, 2001); Violations of Canonical Form that are detected peripherally, are semantically salient and can affect the likelihood of fixating on an item (Becker, Pashler, & Lubin, 2007) and novelty detection i.e. popping out objects becoming less salient over time (Markou & Singh, 2003). The production-level working system will provide a taxonomy of objects in relation to scene schemata as a library limiting the need for further experiments or manual input. We will evaluate the tool by presenting it to skilled experts in game level design as well as integrate and validate our model against documented low-level predictors. We plan to combine our saliency model with a low level saliency model to obtain even more accurate visual attention predictions. Simulating natural effects such as Depth-Of-Field, Camera motion and Dynamic Lighting could benefit from a list of potentially attended objects based on high level saliency. It has been shown that when these effects are dynamically adapted depending on gaze, users reproduce distances better in a VE (Moehring, Gloystein, & Doerner, 2009). We are currently working on a Level-Of-Detail framework for mobile devices that takes into account the dependence of attention deployment on scene context and object topology. The framework accelerates rendering by automatically removing perceptually non-important details in regions that are not expected to be attended (Koulieris, Drettakis, Cunningham, & Mania, 2014b). Taking into account the dependence of attention deployment on scene context and object topology the innovative renderer presented saves computational time by automatically and seamlessly removing perceptually non-important details. We will show that integration of a high level saliency model in a level of detail manager enables the usage of complex effects in lowpower devices by applying them sparingly only in regions that are expected to be attended. Finally, we intend to integrate our model in a game engine for on-the-fly level difficutly adjustments and a smarter game AI. Objects could be repositioned dynamically resulting in an adjustable level of difficulty depending on user performance so far. Object placement could automatically shift after every respawn when a player comes back to life after being killed. A smarter AI could use high level saliency data to spawn opponents that pop-out or appear inconspicuously.

4.2 High Level Saliency Prediction for Level-Of-Detail

This section has been adapted from (Koulieris et al., 2014b)

Attention-based Level-Of-Detail (LOD) managers downgrade the quality of areas that are expected to go unnoticed by an observer to economize on computational resources. The perceptibility of lowered visual fidelity is determined by the accuracy of the attention model that assigns quality levels. Most previous attention based LOD managers do not take into account saliency provoked by context, failing to provide consistently accurate attention predictions. In this work, we extend a recent high level saliency model with four additional components yielding more accurate predictions: an object-intrinsic factor accounting for canonical form of objects, an object-context factor for contextual isolation of objects, a feature uniqueness term that accounts for the number of salient features in an image, and a temporal context that generates recurring fixations for objects inconsistent with the context. We conduct a perceptual experiment to acquire the weighting factors to initialize our model. We design C-LOD, a LOD manager that maintains a constant frame rate on mobile devices by dynamically re-adjusting material quality on secondary visual features of non-attended objects. In a proof of concept study we establish that by incorporating C-LOD, complex effects such as parallax occlusion mapping usually omitted in mobile devices can now be employed, without overloading GPU capability and, at the same time, conserving battery power.

4.2.1 Introduction

LOD algorithms render with higher visual fidelity those regions of a synthetic image that are expected to receive attention, allowing more efficient distribution of the limited resources of a graphics subsystem. The interest in efficient LOD management has been recently renewed due to the explosive growth of the mobile market, which is extremely diverse in terms of computing power. Hardware restrictions of mobile devices prohibit the use of complex effects that demand multiple texture fetches or intense Arithmetic Logic Unit (ALU) operations (Çapin, Pulli, & Akenine-Möller, 2008). An application's expressive power is thus sacrificed in portable devices as content is displayed at degraded levels of detail or quality.

LOD managers have been empowered with perceptual principles in the past to optimize the distribution of computational time and maximize the perceived quality of a rendered scene (D. P. Luebke, 2003). Computation time can be minimized and the quality of an effect downgraded, based on evidence determining that a user is not attending a scene area. A focused distribution of available resources only to *attended* areas allows for higher and more stable frame rates.

Recently, attention assumptions directed by scene context information were utilized to extend a physiologically plausible model of visual attention (M. P. Eckstein, 1998) and use it for game balancing (Koulieris et al., 2014a). The High Level Saliency Model (HLSM) (Koulieris et al., 2014a) incorporates two insights observed in object-context hierarchies. These insights are the *object singletoness hypothesis*, i.e., attention is attracted to physically isolated objects (Theeuwes & Godijn, 2002), and the *scene schema hypothesis* indicating that scenes are comprised of consistent objects expected in a specific context and inconsistent, thus out-of-context objects that are salient (Brewer & Treyens, 1981) (Figure 4.12).

There are, however, other high level saliency phenomena that affect attention which are not captured by this model. We therefore extend the model to include four new factors that allow us to process additional attentional phenomena and predict attention deployment with higher accuracy. In particular we introduce an object-intrinsic factor accounting for the fact that an object pops out if it is rotated in a way that violates its expected posture. The expected posture is known as *canonical form* or canonical orientation (Becker et al., 2007) (Figure 4.12). We also add an *object-context factor* for contextual isolation of objects, a *feature uniqueness term* that accounts for the number of salient features in an image, and a *temporal context* that generates recurring fixations for objects inconsistent with the context or in a non-canonical form.

We then incorporate this new model of attention into a perceptually optimized renderer for mobile platforms that takes into account the dependence of attention deployment on scene context and object placement. This saves computational time by automatically and seamlessly removing perceptually non-important details. Integration of a contextual attention model in a LOD manager enables the usage of – otherwise omitted – complex effects in low-power devices by applying them sparingly only in regions that are expected to be

Figure 4.12: From left to right: A remote control is inconsistent with the sink context. The flowerpot is physically isolated. The tablet is contextually isolated. The chair is in a non-canonical form.

attended and improves battery life by reducing GPU utilization.
We make four primary contributions:

- We extend the High Level Saliency Model (Koulieris et al., 2014a) by introducing the four additional factors described above. These additions yield more accurate predictions of attention than previous work.
- We acquire the parameters to initialize our model's canonical form and perception of contextual singletoness in a perceptual experiment. The experimental design controls for attentional effects from low level features such as luminance or contrast, allowing us to examine the unique contribution of context.
- We develop a novel LOD manager that speeds up rendering in mobile devices based on attention predictions as derived by our model. A proof-of concept implementation selects an appropriate LOD in real-time for subsurface scattering, complex refraction and bump mapping algorithms.
- We demonstrate the accuracy of our implementation by comparing its performance to actual eye-tracking data. We also acquire mobile GPU performance statistics to ensure model effectiveness and quantify battery performance gain when limiting GPU utilization.

4.2.2 Related Work

LOD Modern video games consist of various interconnected software components such as a graphics engine and an audio engine that share hardware resources. LOD methods are essential to improve the interactivity and responsiveness of graphics systems by distributing resources to the image regions that are expected to be attended (D. P. Luebke, 2003). Traditional LOD approaches reduce polygon count by selecting an appropriate instance of polygonal complexity for each model depending on its importance (D. P. Luebke, 2003). Object importance can be determined by attention deployment over the scene or perceptually motivated criteria such as the projected screen size of the object, eccentricity and velocity of objects (Clark, 1976).
Polygonal counts are usually low in mobile devices and mobile GPUs are fill-rate bound deeming polygonal complexity LOD algorithms ineffective (Çapin et al., 2008). Shaders reproduce high quality visual details by exchanging polygonal complexity for additional ALU operations and heavy texture memory accesses. As computation power in mobile GPUs increases faster than memory bandwidth (Owens, 2005) our LOD manager significantly reduces texture fetches.
Attention based LOD Gaze (Loschky & McConkie, 2000) and task (Cater et al., 2003) based LOD managers render the 2 degree fovea region in high quality i.e. the high-resolution part of the visual field and the periphery of vision with less detail. However, LOD management based on gaze encounters difficulties to maintain display updates without artifacts after fast eye saccades. Driving LOD based on pre-defined task areas is limited since it is impossible to quantify the nearly infinite number of potential tasks.
Since low level image features such as luminance, contrast and motion are known to attract attention (Itti, Koch, Niebur, et al., 1998), objects saliency models based on low-level features combined with task relevant information have been employed in order to drive LOD (S. Lee et al., 2009; Hillaire et al., 2010). Since high-level, cognitive phenomena also affect attention, low-level saliency models sometimes fail to predict fixations, especially when an observer manipulates interactive scenes (Sundstedt et al., 2008). Here, we develop and

employ a sophisticated, multi-factor, context-based, attention predictor for interactive environments that takes into account contextual information about a scene to optimize LOD for mobile platforms.

Scene Semantics Object perception in natural scenes relies on the integration of pre-existing knowledge with recently acquired from attentional processing (J. M. Henderson et al., 1999; R. A. Rensink, 2000). In a schema-based LOD framework consistent objects are rendered with lower quality without affecting information uptake (Zotos et al., 2009).

The High Level Saliency Model (HLSM) of (Koulieris et al., 2014a) is based on the Differential-Weighting Model (DWM) that simulates attentional processing using Gaussian combination rules (M. P. Eckstein, 1998). The HLSM takes into account the fact that high level features of an image such as physical object isolation and object-context consistency (Figure 4.12) affect attention. The model describes two *High Level Saliency* sensory units that react to the existence of high level features in the Field-Of-View (FOV). The areas that are most consistent to the feature to which a unit is sensitive capture that unit's attention. For example, an object placed in an unexpected location will elicit a very strong response in a unit that observes object-context consistency. The model predicts saccadic targets by combining the estimated firing rate for both its units in a winner-take-all network.

Each unit's firing behaviour is encoded in Bayesian Priors and is affected by physiological noise having a Gaussian distribution with a mean d'_j and a standard deviation σ (Tolhurst et al., 1983). For each sensory unit j the model calculates for each pixel (x, y) of an image f its probability to be attended as the likelihood ratio LR (Green et al., 1966) of two noisy responses: the likelihood $l_{j,x,y,f}|s$ of observing a feature when it is present and the likelihood $l_{j,x,y,f}|n$ of erroneously observing a feature when it is not present.

$$LR_{j,x,y,f} = \frac{l_{j,x,y,f}(\lambda_{j,x,y,f}|s)}{l_{j,x,y,f}(\lambda_{j,x,y,f}|n)} = exp\left(\frac{\lambda_{j,x,y,f}{d'}_j - 0.5{d'}_j^2}{\sigma^2}\right) \tag{4.9}$$

The posterior probability of each image pixel to be attended is estimated as the weighted average of two units defined, namely $LR_{physical}$ and $LR_{semantic}$ encoding physical object isolation and semantic consistency. The averaging weights were derived from the results of a perceptual study.

Since the plot of many game genres is based on recognizing and acquiring objects in cluttered environments, (Koulieris et al., 2014a) hypothesized and successfully verified that the manipulation of object placement to alter object-context consistency as well as the relocation of physically isolated objects can implicitly adjust game level difficulty. The performance of this model was not evaluated via eye tracking.

4.2.3 Overview

Gaze allocation is influenced by several other *high level factors* in cluttered environments. Not taking these factors into account deprives the model of important contextual information that would otherwise predict attention with higher accuracy. We set three criteria to be satisfied when introducing a high level component in our saliency model. The components (i) should affect attention as documented in cognitive psychology literature, (ii) should be measurable and their parameters quantifiable (iii) should be observed in a video game. We introduce four additional components. First, we subdivide the physically compound state defined in (Koulieris et al., 2014a) by introducing two sub-states based on findings from psychological research (e.g., (Koffka, 1935)). Specifically, we hypothesize that a physically compound object can either be *contextually compound* or *contextually isolated*. Objects belonging in a set are contextually compound. An object positioned in-between a set of similar objects but dissimilar from those in the set, is hypothesized to pop out even when not salient in terms of e.g. color (Koffka, 1935). For example, a tablet computer placed in-between magazines is salient (Figure 4.12). Second, we integrate an object-intrinsic assumption. The three-quarters object view, that makes a large number of surfaces visible is considered to be an object's canonical form (Blanz, Tarr, Bülthoff, & Vetter, 1999; Secord, Lu, Finkelstein, Singh, & Nealen, 2011). The amount of angular deviation from this standard posture affects the object's saliency (Becker et al., 2007) (Figure 4.12). Objects whose orientation is non-canonical are common in games e.g. dead characters or overturned vehicles. Third, we account for object coherence in time. An attended location is usually prevented from being attended again (M. I. Posner & Cohen, 1984), an observation that has been used for LOD management (Longhurst et al., 2006). However, there is strong evidence that recurring fixations are generated for objects that are inconsistent with the context or for objects that are in a non-canonical form

(Becker et al., 2007; J. M. Henderson et al., 1999). Finally, we complement our model by accounting for the biologically motivated feature uniqueness property. A single salient feature in an image pops-out more intensely than when several salient features exist (Itti, Koch, Niebur, et al., 1998; Frintrop et al., 2010b). Our LOD manager adjusts LOD only during player motion. Pop-out artifacts (D. P. Luebke, 2003) are eliminated by exploiting the observer insensitivity to perceive changes occuring during a brief interruption known as the Change Blindness phenomenon (D. J. Simons & Levin, 1997). We evaluate our model via eye-tracking (A. Duchowski, 2007).

4.2.4 Attention Model

We extend the HLSM (Koulieris et al., 2014a) with four additional components: (i) an object-context singletoness factor that traces contextual object isolation, (ii) an object-intrinsic cognitive factor, termed canonical form of objects (Becker et al., 2007), (iii) a biologically motivated feature uniqueness factor (Frintrop et al., 2010b) and (iv) a factor for temporal object coherence. The first two new factors (contextual isolation and canonical form) are incorporated through two new high level sensory units. To account for feature uniqueness, equations determine the number of local maxima found in the probability output of a sensory unit. That is, the more maxima there are, the less unique a feature is. For example, if there is only a single violation of canonical form, its uniqueness weight is high. If several violations exist, all violations are less unique. Recurring fixations to areas containing canonical form violations or schema inconsistencies are generated by multiplying a unit's current output with a number of logarithmically attenuated previous outputs.
The output of all units is multiplied with a feature uniqueness weight:

$$w_{unit}^{unq} = \frac{1}{|\vee|P_{unit,x,y,f}} \tag{4.10}$$

x, y denotes image location, f denotes frame number, $|\vee|$ the number of posterior probability local maxima (Figure 4.13).

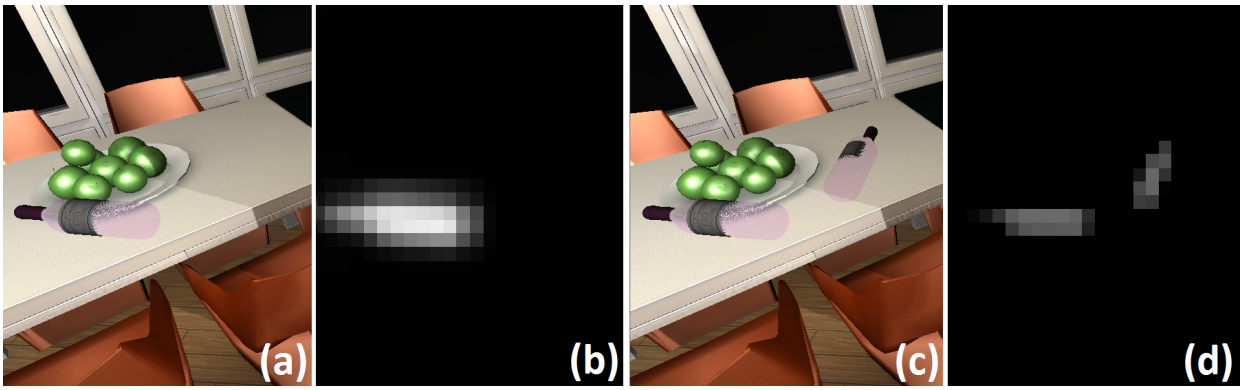


Figure 4.13: A single violation of canonical form in the FOV (a) provokes a response in the canonical form sensory unit (b). When more violations of canonical form exist (c) the sensory unit's output is attenuated (d).

The output of the schema consistency unit and the canonical form unit (Equation 1) are also multiplied with a temporal context weight:

$$w_{unit,x,y,f}^{tmp} = \prod_{f=1}^{F} P_{unit,x,y,f} e^{-af} \tag{4.11}$$

F the number of previous frames examined, a is a user-defined attenuation factor (Figure 4.14).
The posterior probability $\mathbf{P}_{x,y,f}$ that an observer attends an image location, as part of our enhanced model, is linearly estimated (Frintrop et al., 2010b) from both the semantic consistency *(sem)* and physical isolation *(phy)* units of (Koulieris et al., 2014a) combined with our contextual isolation *(cnt)* and canonical form *(cfr)* units, updated for feature uniqueness and temporal context:

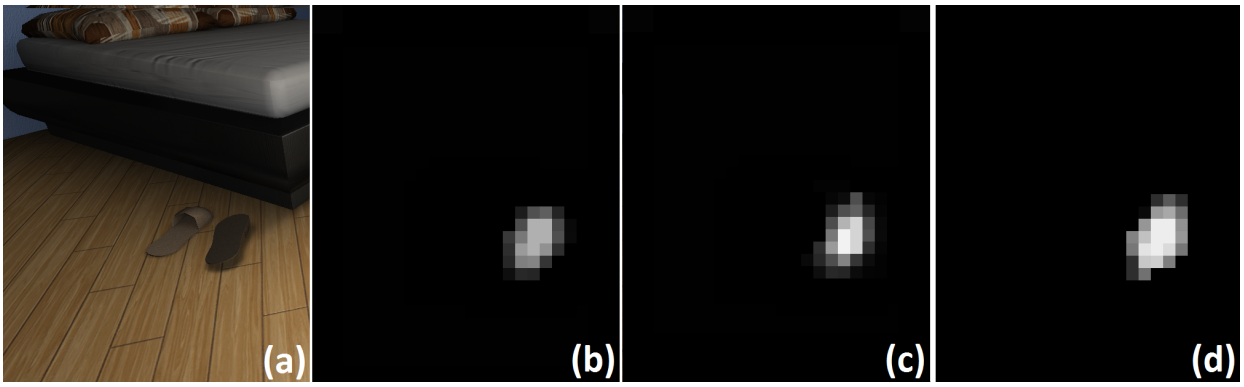


Figure 4.14: The slipper on the right is in a non-canonical form (a). The output of the canonical form unit is shown in the current frame (b), in a subsequent frame (c) and in a third frame after the first (d). The increasing probability will generate recurring fixations for our model.

$$\mathbf{P}_{x,y,f} = w_{sem} w^{unq}_{sem} w^{tmp}_{sem,x,y,f} \mathbf{P}_{\mathbf{sem},x,y,f} + w_{phy} w^{unq}_{phy} \mathbf{P}_{\mathbf{phy},x,y,f} + w_{cnt} w^{unq}_{cnt} \mathbf{P}_{\mathbf{cnt},x,y,f} + w_{cfr} w^{unq}_{cfr} w^{tmp}_{cfr,x,y,f} \mathbf{P}_{\mathbf{cfr},x,y,f} \quad (4.12)$$

In Section 5 the contribution weights w_{sem} and w_{phy} that were estimated in (Koulieris et al., 2014a) are adapted to our model and the weights w_{cnt} and w_{cfr} are estimated based on our experimental data.

4.2.5 Perceptual Study

We conducted a perceptual experiment using a *Search* task to be comparable to (Koulieris et al., 2014a). We thus: (i) examine the effect of violations of canonical form and contextual singletoness on visual attention and (ii) obtain contribution weights of each factor for our model.
Stimuli We factorially combined the two factors to control the spatial arrangement of three objects (a tablet computer, a pair of spectacles and a remote control; see Figure 4.15) in four virtual environments. The four scenes were contextually compound/canonical, contextually compound/non-canonical, contextually singleton/canonical, or contextually singleton/non-canonical (Figure 4.16). All objects were consistent with the scenes and were physically compound. The Saliency Toolbox (Walther & Koch, 2006) (Figure **??**) was used to ensure that the three objects had a minimum low-level saliency.

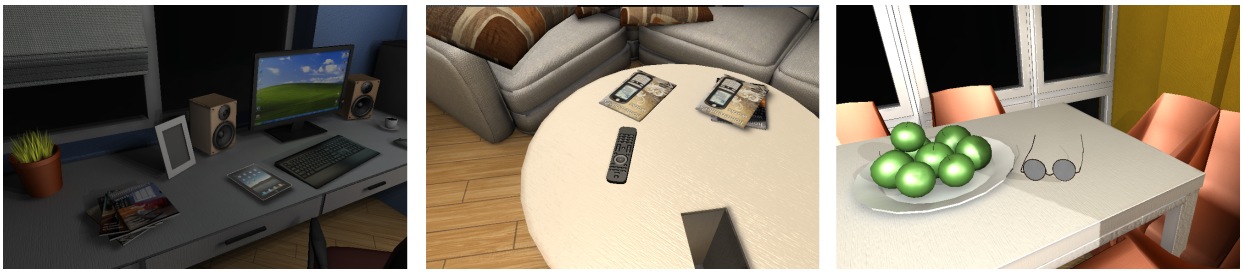


Figure 4.15: The subjects searched for three objects, a tablet computer, a remote control and a pair of spectacles.

Participants Fourty-eight people participated (8 female, mean age 23) in the experiment, with 12 people being assigned to each of the 4 conditions.
Apparatus The stimuli were displayed on a NVisor^{TM} SX111 Head Mounted Display (HMD), which has stereo SXGA resolution and a FOV of 102 degrees horizontal by 64 degrees vertical. Participants moved

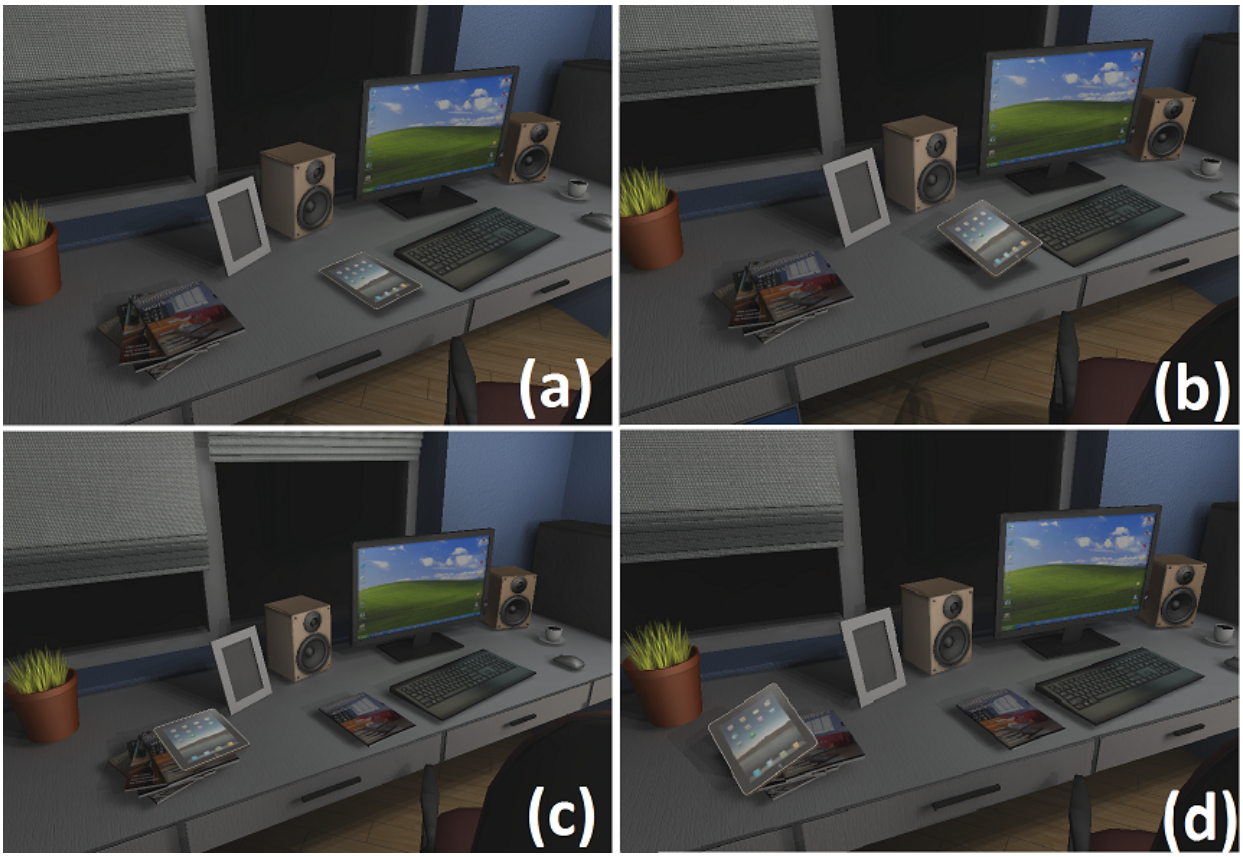


Figure 4.16: The tablet is in a (a) contextually compound canonical form (tablet and keyboard), (b) contextually compound non-canonical form (slanted, (c) contextually singleton canonical form and (d) contextually singleton non-canonical form.

through the virtual environment using a game-pad for translation and an InterSense^{TM} InertiaCube3^{TM} 3DoF head tracker for rotation. Navigation was restricted to -70/70 degrees vertically. Eye tracking information was recorded using a twin-CCD binocular eye-tracker by Arrington Research^{TM}, which was attached to the HMD. The eye tracker was updated at a frequency of 30Hz.

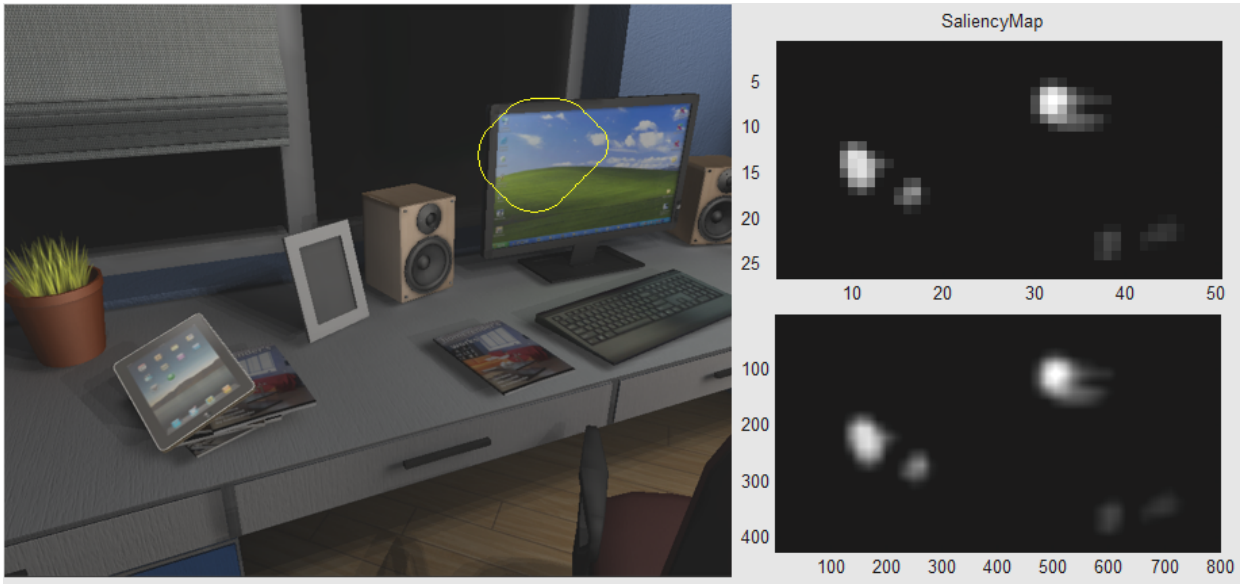
Procedure Participants sat on a swivel chair and were familiarized with the setup in a training session. They were then requested to navigate around the scene in order to find and collect all three objects. Task accuracy, completion time, and eye-tracking data were recorded.

Results Task accuracy was always 100%. On average, participants needed 167.788, 255.386, 82.189, and 195.985 seconds for the compound/canonical, compound/non-canonical, singleton/canonical, and singleton/non-canonical conditions, respectively. Task completion times were analyzed with a linear Hierarchical Multiple Regression analysis (HMR) with contextual singletoness being entered at stage one and canonical form at stage two. HMR fits a linear model to the data, with one term for each factor. The weight associated with each term is related to the correlation coefficient between the dependent variable (here, completion time) and the different factors. This effectively describes how well changes in the measured data can be explained or predicted by changes in the factors. Contextual singletoness contributed significantly to the regression model, $F(1, 46) = 16.83, p < .001$ and accounted for 26.79% of the variation in task completion time. Introducing canonical form explained an additional 51.68%, $F(2, 45) = 82.03, p < .001$, for a total explained variance of 78.47%. The coefficients for the two factors can be seen in Table 4.3. Predictions for a condition can be obtained by combining the intercept (i.e., performance in the Compound/Canonical condition) with the appropriate modifiers (i.e., the non-canonical form and/or singleton terms; see Table 4.3). The predictions of the model are consistent with the actual recorded completion times.

An analysis of the eye-tracking Regions-Of-Interest (ROIs) showed that attention is indeed attracted both

Coefficients	Estimate Time	p-value
Intercept	161.238	< 0.0001
+Non-Canonical Form term	100.697	< 0.0001
+Singleton Placement term	-72.500	< 0.0001

Table 4.3: The regressions coefficients for each factor.



to contextually singleton objects and to objects in a non-canonical form.

Discussion The canonical form and contextual isolation of objects play a significant role in attention deployment. In particular, in the non-canonical form conditions the objects were actively observed despite the fact that their recognition was extremely slow when compared to the canonical form condition. This is apparently in contradiction with (Koulieris et al., 2014a) who found that actively attended salient objects are easy to find. Thus, when managing LOD, an object in non-canonical form is salient and should always be rendered in high quality.

Weight Generation In their paper (Koulieris et al., 2014a) derived the model weights from the correlation coefficients by dividing the amount of variance that a factor explained by the total explained variance. Since a single, between-participants experiment using a factorial combination of all levels of all four factors does not exist (it would require a prohibitively large number of participants), it is not possible to determine the *relative* amount of variance each factor explains. Fortunately, there is an alternative: the regression coefficients explicitly correlate changes in a factor with changes in completion time. Thus, it should be possible to get similar weights directly from the completion times. To make the completion times in two experiments comparable, we use the single condition that is the same in both experiments (physically compound/consistent in (Koulieris et al., 2014a) and contextually compound/canonical here) to normalize the weights. According to the regression models, the completion time should be 140.581 seconds in (Koulieris et al., 2014a) and 161.238 here. Therefore, we divide the actual completion times in all of (Koulieris et al., 2014a)'s conditions by 140.581, and all of our times by 161.238. For example, the actual mean of the contextually compound/canonical condition was 167.788. After normalization, it is 1.04. By comparing the relevant conditions, we can determine the relative effect of altering one factor. For example, we compare changes in contextual isolation for canonical objects ($1.04 - 0.51$) and for non-canonical objects ($1.58 - 1.22$). The average difference is 0.45. After we obtain all weights, we ensure that the weights all sum to 1. The final weights are 0.07 for schema, 0.33 for physical isolation, 0.35 for canonical form and 0.25 for contextual isolation (see supplemental document for more details).

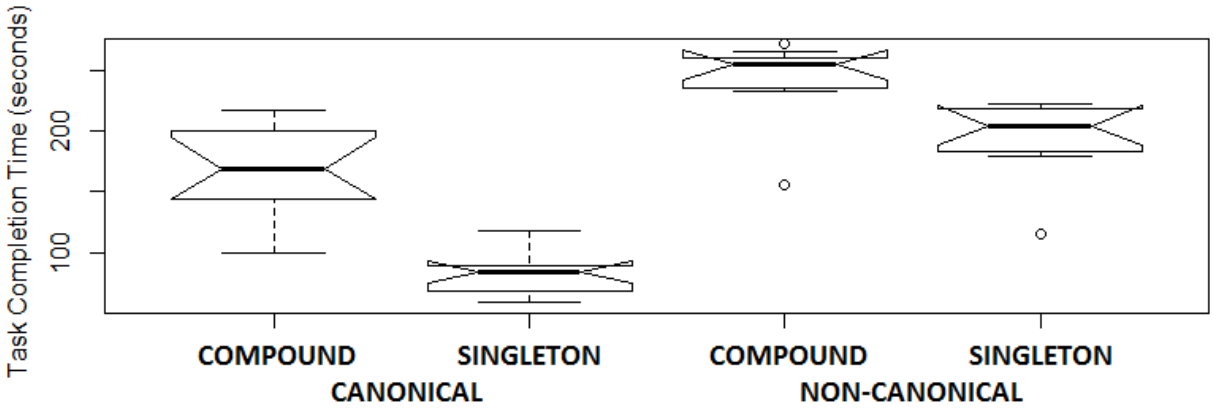


Figure 4.17: Task completion time distribution of the experimental conditions. The median value for each condition is depicted by the horizontal line. The notched boxes depict the middle quartiles. The outer bars represent the extremes for each case.

Figure 4.18: Left to right: Subsurface scattering, refraction and bump mapping low to high quality.

4.2.6 LOD for Mobile Graphics

We developed a generic material LOD manager based on attention for Unity 3D™ game engine that we call C-LOD. C-LOD is a reactive fixed frame rate scheduler (D. P. Luebke, 2003) that constantly examines frame rate and attention deployment predictions using the criteria of our model. When frame rate drops below 30 frames per second on fill-rate bound mobile devices, C-LOD automatically lowers the rendering quality of objects predicted not to be attended until performance is restored (Figure 4.19). The highest quality possible is maintained for all attended objects.

C-LOD Effects

C-LOD can manage any effect that has at least two levels of detail. For this proof-of-concept implementation we selected three complex effects that are usually omitted in mobile devices as they require many texture fetches (Çapin et al., 2008). We used two LOD fall-backs for each effect, that require fewer texture fetches (Figure 4.18).

Subsurface light transport in translucent materials requires intense analytical calculations, making it impossible for mobile devices to render this effect (Jensen, Marschner, Levoy, & Hanrahan, 2001). To simulate the high quality effect, we approximated light transport using a pre-computed map of local thickness for each model calculated by inverting the normals of the model and estimating ambient occlusion with the inverted normals (Barre-Brisebois, 2011). The medium LOD level substitutes the thickness map with a standard distance-attenuated diffuse lighting combined with the distance-attenuated dot product of the view vector and the inverted light vector. The low quality fall-back is an opaque Blinn-Phong specular shader.

Refraction is a computationally expensive effect for mobile devices. OpenGL ES2.0 devices do not support Multiple Render Targets (MRTs) thus existing methods that estimate refraction for both the front and back interfaces of an object are slow (Wyman, 2005). Single interface refraction produces convincing results. Single interface refraction with chromatic aberration (Lindholm, Kilgard, & Moreton, 2001) was selected as the high level refraction effect. The medium effect removes chromatic aberration by exchanging the

wavelength-dependent sampling of the RGB channels with a single lookup, significantly reducing texture fetches by a factor of three. The low quality effect is a uniformly distorted transparent shader.
Bump Mapping via tessellation and displacement mapping is not available on OpenGL ES2.0 devices. For high quality bump mapping we incorporated the texture-heavy Parallax Occlusion Mapping method (Tatarchuk, 2006). For the medium quality level effect, we employed simple parallax mapping that does not support self-shadowing (Kaneko et al., 2001). The low quality is a standard normal mapped shader.

C-LOD Components

The Predictor We implemented our model in the GPU. Our system detects non-canonical object forms by examining object position in relation to the view vector. We utilize object IDs to locate contextually singleton objects. An analytical determination of feature uniqueness would require the calculation of the bi-variate partial derivative of each unit's output. Identifying local maxima in a Gaussian pyramid (Ziegler, Tevs, Theobalt, & Seidel, 2006) is slow on mobile as it uses render buffer ping ponging. We count local maxima by employing an approximation that exploits hardware's linear interpolation capabilities. We render each unit's output in a 4x4 resolution frame buffer object only once each second. By thresholding 16 texel fetches per unit buffer we count up to 16 local maxima competently. We approximate temporal context calculations by storing up to F low resolution previous frame buffer objects and combine them using hardware blending and an 1D ramp texture storing the pre-calculated logarithmically attenuated function (Equation 3). For the scene schemata and physical singleton factors we re-implemented the detectors of (Koulieris et al., 2014a), as well as our extensions described in previous sections. We initialized our model equations using the the weights estimated in Section 5.
The Texel Engine C-LOD's Texel Engine constantly monitors object predictions derived from our attention model. A special 2D texture is updated that works as a material quality lookup table (Figure 4.19). The columns of the texture correspond to all object/material combinations found in a scene and each row represents a LOD for all object/material combinations. A higher row number signifies a more aggressive simplification overall. Introducing a simplification for object/material combination x in row y imposes that all subsequent rows have the same or lower quality for x. This restriction maintains visual coherence between LODs and induces the smallest possible number of quality reductions. As a result, values over the diagonal of the texture are always white signifying the highest quality possible. The system updates the texture once per second in synchronization with camera movement.

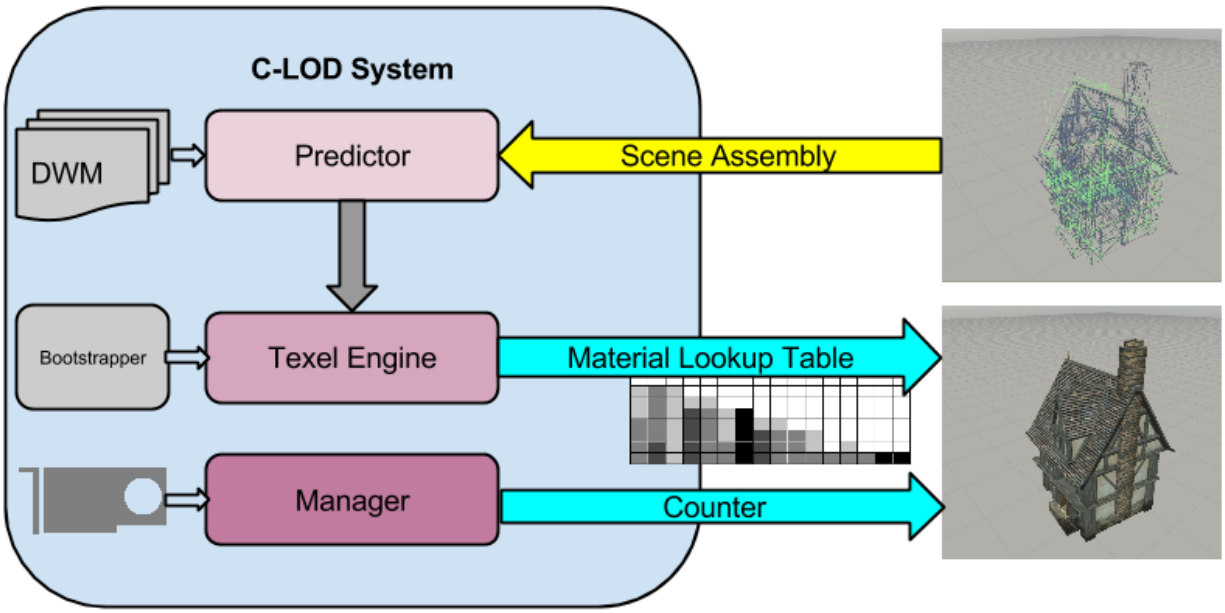


Figure 4.19: The C-LOD system architecture.

Est.	Object gazed	HQ	C-LOD	Total
R	random object	¡ 5%	¡ 5%	¡ 5%
$E1$	1st prediction	40%	42.3%	41.1%
$E2$	1st or 2nd	69.9%	74.8%	72.3%
$E3$	1st or 2nd or 3d	86.9%	92.7%	89.7%

Table 4.4: The ratio of frames that the attended object was predicted correctly for the high quality condition, the C-LOD managed condition and in total. E1 denotes that the gazed object matches the first prediction. E2 denotes that the gazed object matches either the first or the second predicted object. E3 denotes that the gazed object matches either the first, or the second or the third object.

The Bootstrapper The interaction between the graphics processor, CPU and memory of a mobile device is not trivial. When bootstrapping, C-LOD performs system profiling. The materials managed are initially rendered at their lowest quality. Then, in rapid succession, the quality level of each object's material is increased while frame rate is monitored. This procedure determines a scale factor that controls the aggressiveness of simplifications by the Texel Engine.
The Manager A Finite State Machine (FSM) monitors frame rate during execution. When frame rate drops and motion is detected, a counter is increased. This counter is communicated to all managed materials. A re-mapped object ID of each object is appointed as the u coordinate to sample the look-up table texture and the counter variable as the v coordinate. The sampled value is communicated to the fragment shader where it controls a conditional branch that selects the appropriate LOD for the shader or acts as an iteration counter e.g., for ray marching in the parallax occlusion mapping shader. Updating the counter only when camera moves, reduces luminance offsets and flickering effects. Frame rate is constantly re-evaluated and the counter is increased/decreased to maintain the best LOD for the current conditions (Figure 4.19).

4.2.7 Evaluation of C-LOD

We evaluated C-LOD's efficacy both via eye tracking and by acquiring GPU performance data on a mobile device. We also measured battery performance improvement.
Model Accuracy To measure the model's accuracy in predicting attention we performed an experiment on our eye-tracked HMD set-up.
Design To empirically verify that changes in LOD were not perceived and did not affect attention deployment, we rendered a scene consisting of 50k triangles and complex shaders twice. In the first version of the scene (HQ), all effects were set in the highest quality possible. In the second condition (C-LOD) quality was managed by our system. The rendering was performed on a high-end desktop computer to eliminate fluctuations in the frame rate that would have occured in a tablet device inadvertently affecting attention deployment. The FOV of the HMD was restricted to 40 degrees horizontally and 23 degrees vertically to simulate a 10.1" tablet held at a 30cm observer distance (Slater, Spanlang, & Corominas, 2010). Participants were requested to find and collect seven objects placed in consistent, inconsistent, physically isolated, contextually compound, contextually isolated locations and in a canonical/non-canonical form. In total, 22 people participated (2 female, mean age 22), with 11 people in each of the two conditions.
Results In total, $88,404$ object fixations were recorded for all participants (Figure 4.20). Given that human attention may be directed at multiple foci (Awh & Pashler, 2000), we recorded the three most prominent objects predicted to be fixated by our system for each frame of the simulation. We defined three quantitative estimators to denote the ratio of frames that gaze was allocated in an increasingly larger subset of the predicted objects, to the total number of simulation frames. A baseline R estimator was defined that selects a random object in the FOV for each frame. Both conditions yielded similar results. We summarize the estimators and their results in Table 4.4. In short, the addition of the C-LOD changes did not alter gaze performance, and thus were most likely not perceived by the participants.
Model Efficiency To assess the impact of C-LOD on GPU performance we reconstructed $2,947$ seconds of player motion of both experimental conditions on an Android quad-core Cortex A9 1.6GHz OpenGL ES2.0 mobile device and sampled the framerate at a 5Hz rate. A total of $17,681$ frame rate samples were collected. An independent-samples t-test was conducted, revealing a significant difference between the HQ

($M = 24.05, SD = 2.92$) and C-LOD ($M = 25.6, SD = 1.33$) conditions; $t(8, 418) = -44.16, p < 0.0001$. The C-LOD condition exhibits a consistently stabler frame rate and provides a slightly higher mean frame rate when compared to the HQ quality setting (Figure 4.21). The Android Debug Bridge (ADB) and Tracer for OpenGL tools were employed to conduct a deep frame inspection. C-LOD estimations run for 4ms on average per frame. Given the increase in mean frame rate between the two conditions it can be concluded that this cost is amortized between frames.

Battery life improvement Quering ADB indicated that the battery's average voltage drop was 21mVolts greater for the HQ condition versus the C-LOD managed condition. This indicates an increased discharge rate that was also portrayed in the total run time. Player motion data from the validation experiment were re-played in the HQ and C-LOD settings until battery run out. The C-LOD condition lasted 249 minutes; the HQ condition lasted for 233 minutes.

Discussion Results indicate that C-LOD identifies the observed object 8 times better than a random estimator in the worst case (Table 4.4). For three attended objects prediction rate approaches 90%. This suggests that quality reductions go mostly unnoticed. Integrating C-LOD in a mobile 3D graphics application stabilizes frame rate without sacrificing perceived quality and boosts battery run time by 6.5% (Figure 4.21).

4.2.8 Conclusion

We presented an extension to the HLSM (Koulieris et al., 2014a) by introducing four novel factors that affect attention deployment: object canonical form, contextual singletoness, feature uniqueness and temporal context. We acquired the parameters to initialize our model in a perceptual experiment. We developed a LOD manager for mobile devices that maintains a constant framerate by selecting an appropriate LOD for materials based on attention. We evaluate the performance our algorithm via eye-tracking and by acquiring GPU performance data on mobile devices, confirming that complex effects such as parallax occlusion mapping that are usually omitted in mobile devices can now be employed without exhausting GPU capability. We verified an increase in battery life due to less GPU utilization.

Future work includes extending our model with low-level factors for more accurate predictions when gross low level irregularities exist in an image. We will investigate the performance of the proposed LOD manager in dynamic scenes. An attention based cinematography system could be developed that applies post-process effects such as Depth-Of-Field based on attention.

Acknowledgments This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II: Investing in knowledge society through the European Social Fund. We thank Adobe and Autodesk for generous donations; the work was partly supported by EU FP7 project ICT-611089-CR-PLAY www.cr-play.eu.

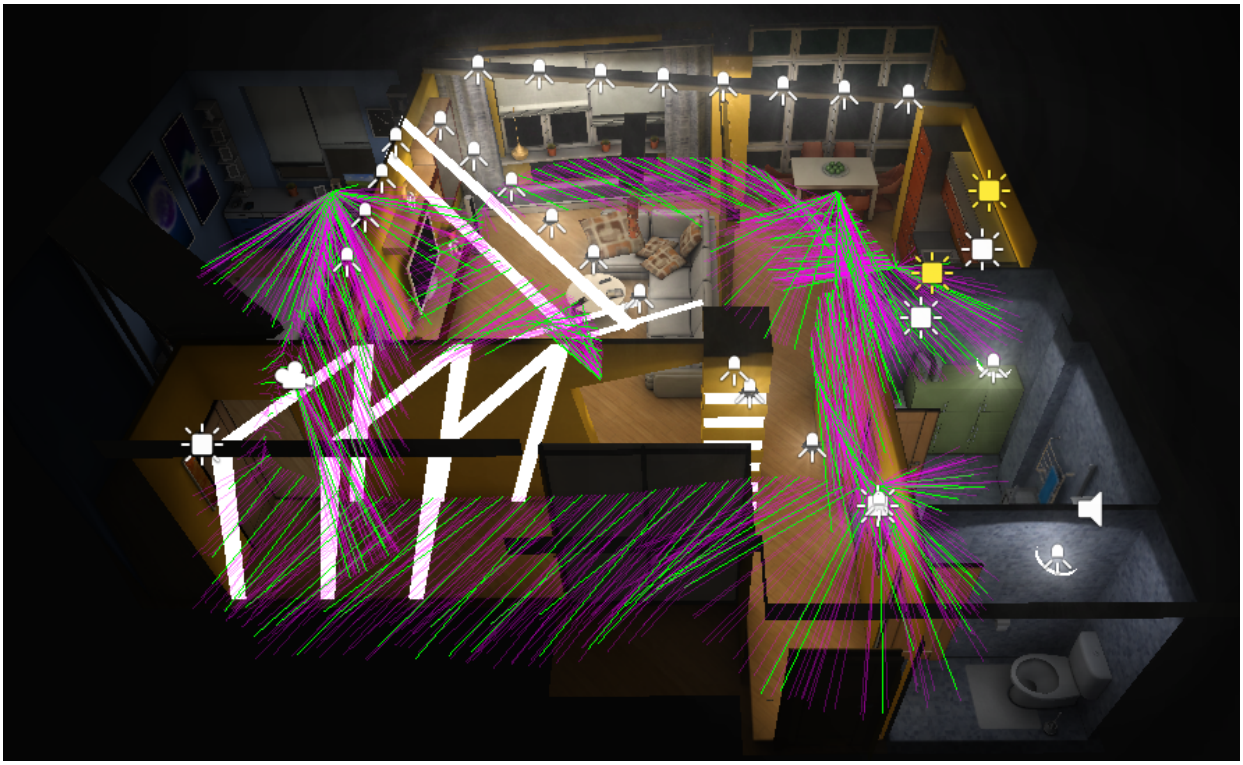


Figure 4.20: Our validation tool indicates the subject's gaze point with magenta colored beams. The green beams indicate predictions by our attention model.

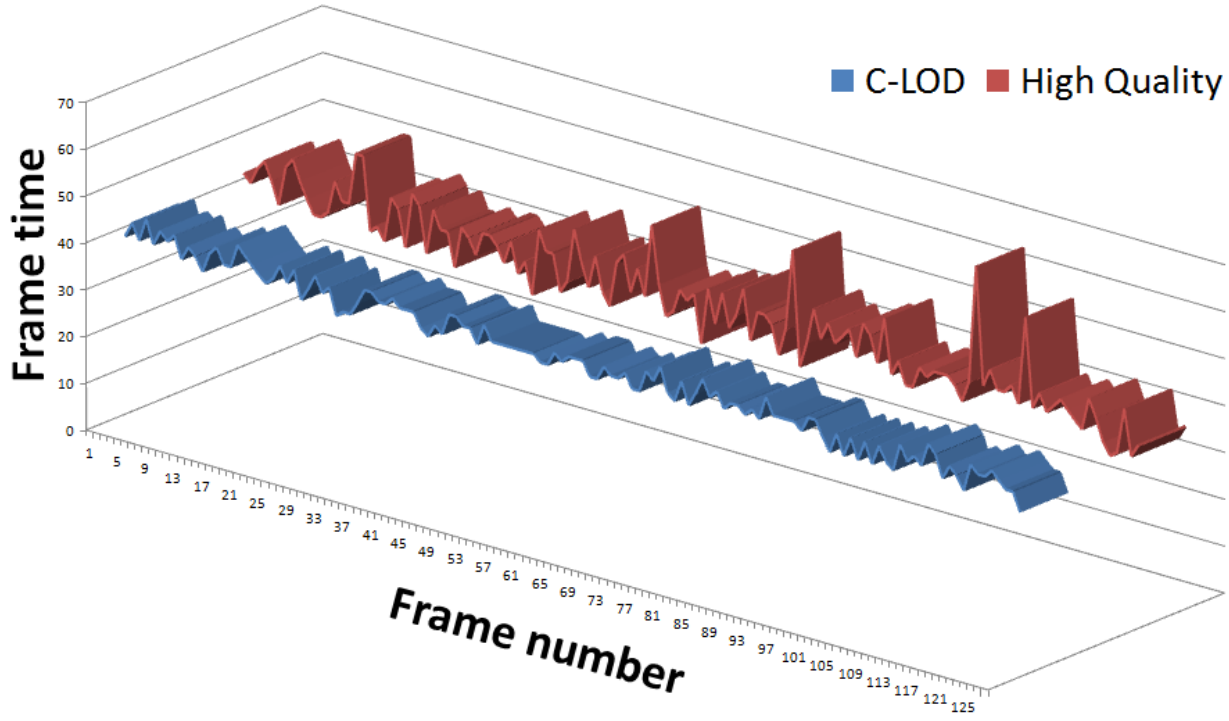


Figure 4.21: Frame time for 128 random sequential frames of the HQ and C-LOD conditions. Notice the intense fluctuation of the frame time in the HQ condition when compared to the C-LOD condition.

References

Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1597–1604).

Agresti, A. (1989, Oct). A survey of models for repeated ordered categorical response data. *Stat Med*, *8*(10), 1209-24.

Ajallooeian, M., Borji, A., Araabi, B., Ahmadabadi, M., & Moradi, H. (2009). Fast hand gesture recognition based on saliency maps: An application to interactive robotic marionette playing. In *Robot and human interactive communication, 2009. ro-man 2009. the 18th ieee international symposium on* (pp. 841–847).

Akamine, K., Fukuchi, K., Kimura, A., & Takagi, S. (2012). Fully automatic extraction of salient objects from videos in near real time. *The Computer Journal*, *55*(1), 3–14.

Allport, A., Meyer, D., & Kornblum, S. (1993). Attention and control: Have we been asking the wrong questions? a critical review of twenty-five years. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience, MIT Press, Cambridge, Mass*, 183–218.

American Cancer Society. (2014). *Cancer Facts & Figures.* (http://www.cancer.org/ Research/CancerFactsFigures/index)

Andolina, S., Santangelo, A., Cannella, M., Gentile, A., Agnello, F., & Villa, B. (2009). Multimodal virtual navigation of a cultural heritage site: the medieval ceiling of steri in palermo. In *Proceedings of the 2nd conference on human system interactions* (pp. 559–564). Piscataway, NJ, USA: IEEE Press. Retrieved from `http://portal.acm.org/citation.cfm?id=1689359.1689455`

AO, B.-W., Jr, F. C., LW, N., & JY, L. (2003). Application of likelihood ratio to classification of mammographic masses; performance comparison to case-based reasoning. *Medical Physics*, *30*, 949–958.

Arbib, M., & Didday, R. (1971). The organization of action-oriented memory for a perceiving system. part i: The basic model. *Journal of Cybernetics*, *1*(1), 3-18.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, *12*(2), 157–162.

Avraham, T., & Lindenbaum, M. (2010). Esaliency (extended saliency): Meaningful attention using stochastic image modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(4), 693–708.

Awh, E., Belopolsky, A., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends in Cognitive Sciences*.

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 834.

Aziz, M., & Mertsching, B. (2008). Fast and robust generation of feature maps for region-based visual attention. *Image Processing, IEEE Transactions on*, *17*(5), 633–644.

Bahrami, B., Lavie, N., & Rees, G. (2007, March). Attentional load modulates responses of human primary visual cortex to invisible stimuli. *Current Biology*, *17*(6), 509-513.

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2007). Subtle gaze direction. In *Siggraph '07: Acm siggraph 2007 sketches* (p. 44). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/1278780.1278834

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2008). Subtle gaze direction. *ACM Transactions on Graphics*. (To appear)

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2009a). Subtle gaze direction. *ACM Transactions on Graphics (TOG)*, *28*(4), 100.

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2009b, September). Subtle gaze direction. *ACM Trans. Graph.*, *28*, 100:1–100:14. Retrieved from `http://doi.acm.org/10.1145/1559755.1559757` doi: http://doi.acm.org/10.1145/1559755.1559757

Bailey, R., McNamara, A., Sudarsanam, N., & Grimm, C. (2009c, September). Subtle gaze direction. *ACM Trans. Graph.*, *28*, 100:1–100:14. Retrieved from `http://doi.acm.org/10.1145/1559755.1559757` doi: http://doi.acm.org/10.1145/1559755.1559757

Baldi, P., & Itti, L. (2010). Of bits and wows: a bayesian theory of surprise with applications to attention. *Neural Networks*, *23*(5), 649–666.

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Baluch, F., & Itti, L. (2011, March). Mechanisms of top-down attention [td;bu;psy;mod;fmri;eye;phy]. *Trends in Neurosciences*, *34*, 210-224. (Front cover of journal, April 2011; Ranked No. 1 TiNS most downloaded articles in 2011.)

Ban, S., Kim, B., & Lee, M. (2010). Top-down visual selective attention model combined with bottom-up saliency map for incremental object perception. In *Neural networks (ijcnn), the 2010 international joint conference on* (pp. 1–8).

Ban, S., Lee, I., & Lee, M. (2008). Dynamic visual selective attention model. *Neurocomputing*, *71*(4), 853–856.

Bar, M., Ullman, S., et al. (1996). Spatial context in recognition. *PERCEPTION-LONDON-*, *25*, 343–352.

Barre-Brisebois, C. (2011). Approximating translucency for a fast, cheap and convincing subsurface-scattering look. In *Game developers conference.*

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Baudisch, P., DeCarlo, D., Duchowski, A., & Geisler, W. (2003). Focusing on the essential: considering attention in display design. *Commun. ACM*, *46*(3), 60–66. doi: http://doi.acm.org/10.1145/636772.636799

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 20.

Belardinelli, A., Pirri, F., & Carbone, A. (2006). Robot task-driven attention. In *Proceedings of the 2006 international symposium on practical cognitive agents and robots* (pp. 117–128).

Bernhard, M., Stavrakis, E., & Wimmer, M. (2010). An empirical pipeline to derive gaze prediction heuristics for 3d action games. *ACM Transactions on Applied Perception (TAP)*, *8*(1), 4.

Bernhard, M., Zhang, L., & Wimmer, M. (2011). Manipulating attention in computer games. In *Ivmsp workshop, 2011 ieee 10th* (pp. 153–158).

Beuter, N., Lohmann, O., Schmidt, J., & Kummert, F. (2009). Directed attention-a cognitive vision system for a mobile robot. In *Robot and human interactive communication, 2009. ro-man 2009. the 18th ieee international symposium on* (pp. 854–860).

Bian, P., & Zhang, L. (2009). Biological plausibility of spectral domain approach for spatiotemporal visual saliency. *Advances in Neuro-Information Processing*, 251–258.

Blanz, V., Tarr, M. J., Bülthoff, H. H., & Vetter, T. (1999). What object attributes determine canonical views? *Perception-London*, *28*(5), 575–600.

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

Bonev, B., Chuang, L., & Escolano, F. (2012). How do image complexity, task demands and looking biases influence human gaze behavior? *Pattern Recognition Letters*.

Bordwell, D. (2011). *http : //www.davidbordwell.net/blog/*2011/02/14/*watching − you − watch − there − will − be − blood/*.

Borji, A. (2010). *https://sites.google.com/site/saliencyevaluation/*.

Borji, A. (2012). Boosting bottom-up and top-down visual features for saliency estimation. In *2012 ieee conference on computer vision and pattern recognition* (pp. 438–445).

Borji, A., Ahmadabadi, M., & Araabi, B. (2011). Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications*, *22*(1), 61–76.

Borji, A., Ahmadabadi, M., Araabi, B., & Hamidi, M. (2010). Online learning of task-driven object-based visual attention control. *Image and Vision Computing*, *28*(7), 1130–1145.

Borji, A., & Itti, L. (2012, Jun). Exploiting local and global patch rarities for saliency detection [bu;mod]. In *Proc. ieee conference on computer vision and pattern recognition (cvpr), providence, rhode island* (p. 1-8).

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1).

Borji, A., Sihite, D. N., & Itti, L. (2012a, Aug). An object-based bayesian framework for top-down visual attention [bu;td;mod;cv]. In *Proc. twenty-sixth aaai conference on artificial intelligence (aaai-12), toronto, canada* (p. 1529-1535).

Borji, A., Sihite, D. N., & Itti, L. (2012b). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study [mod;bu]. *IEEE Transactions on Image Processing*, 1-16 (in press).

Borji, A., Sihite, D. N., & Itti, L. (2012c). What/where to look next? modeling top-down visual attention in complex interactive environments [mod;td]. *IEEE Transactions on Systems, Man, and Cybernetics, Part A - Systems and Humans*, 1-16 (in press).

Boscotrecase. (1BC). *Landscape with perseus and andromeda , http://www.metmuseum.org/toah/works-of-art/20.192.16.*

Brefczynski, J. A., & DeYoe, E. A. (1999). A physiological correlate of the 'spotlight' of visual attention. *Nat Neurosci*, *2*(4), 370-374.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*(2), 207–230.

Bruce, N. D. B., & Tsotsos, J. K. (2005). Saliency based on information maximization. In *Nips*.

Bruns, E., Brombach, B., Zeidler, T., & Bimber, O. (2007, April). Enabling mobile phones to support large-scale museum guidance. *IEEE MultiMedia*, *14*, 16–25. Retrieved from `http://portal.acm.org/citation.cfm?id=1262178.1262247` doi: 10.1109/MMUL.2007.33

Burattini, E., Rossi, S., Finzi, A., & Staffa, M. (2010). Attentional modulation of mutually dependent behaviors. *From Animals to Animats 11*, 283–292.

Butko, N., & Movellan, J. (2009). Optimal scanning for faster object detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 2751–2758).

Çapin, T. K., Pulli, K., & Akenine-Möller, T. (2008). The state of the art in mobile graphics research. *IEEE Computer Graphics and Applications*, *28*(4), 74–84.

Carmi, R., & Itti, L. (2006, Aug). The role of memory in guiding attention during natural vision [mod;bu;td;eye]. *Journal of Vision*, *6*(9), 898-914.

Carmody, D. P., Kundel, H. L., & Toto, L. C. (1984). Comparison scans while reading chest images. Taught, but not practiced. *Invest Radiol*, *19*, 462-466.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1980). An analysis of perceptual and cognitive factors in radiographic interpretation. *Perception*, *9*, 339-344.

Carrasco, M. (2011). Visual attention. *Vision research*.

Cater, K., Chalmers, A., & Ledda, P. (2002). Selective quality rendering by exploiting human inattentional blindness: looking but not seeing. In *Vrst '02: Proceedings of the acm symposium on virtual reality software and technology* (pp. 17–24). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/585740.585744

Cater, K., Chalmers, A., & Ward, G. (2003). Detail to attention: exploiting visual tasks for selective rendering. In *Proceedings of the 14th eurographics workshop on rendering* (pp. 270–280).

Cave, K. R. (1999). The featuregate model of visual selection. *Psychological research*, *62*(2), 182–194.

Cerf, M., Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12).

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, *20*.

Chen, D. R., Chang, R. F., & Huang, Y. L. (1999, Nov). Computer-aided diagnosis applied to US of solid breast nodules by using neural networks. *Radiology*, *213*, 407–412.

Chentanez, N., Alterovitz, R., Ritchie, D., Cho, L., Hauser, K. K., Goldberg, K., ... O'Brien, J. F. (2009, Aug). Interactive simulation of surgical needle insertion and steering. In *Proceedings of acm siggraph 2009* (p. 88:1-10). Retrieved from `http://graphics.berkeley.edu/papers/Chentanez-ISN-2009-08/` doi: 10.1145/1576246.1531394

Chikkerur, S., Serre, T., Tan, C., & Poggio, T. (2010). What and where: A bayesian inference theory of attention. *Vision research*, *50*(22), 2233–2247.

Chou, S.-C., Hsieh, W.-T., Gandon, F. L., & Sadeh, N. M. (2005). Semantic web technologies for context-aware museum tour guide applications. In *Proceedings of the 19th international conference on advanced information networking and applications - volume 2* (pp. 709–714). Washington, DC, USA: IEEE Computer Society. Retrieved from `http://dx.doi.org/10.1109/AINA.2005.307` doi: http://dx.doi.org/10.1109/AINA.2005.307

Choudary, O., Charvillat, V., Grigoras, R., & Gurdjos, P. (2009). March: mobile augmented reality for cultural heritage. In *Proceedings of the 17th acm international conference on multimedia* (pp. 1023–1024). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1631272.1631500` doi: http://doi.acm.org/10.1145/1631272.1631500

Chun, M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology.*, *36*, 28–71.

Clark, J. H. (1976). Hierarchical geometric models for visible surface algorithms. *Communications of the ACM*, *19*(10), 547–554.

Cole, F., DeCarlo, D., Finkelstein, A., Kin, K., Morley, K., & Santella, A. (2006, June). Directing gaze in 3D models with stylized focus. *Eurographics Symposium on Rendering*, 377–387.

Crick, F. (1984). Function of the thalamic reticular complex: the searchlight hypothesis. *Proc Natl Acad Sci USA*, *81*, 4586-4590.

Cunningham, D. W., & Wallraven, C. (2011). *Experimental design: From user studies to psychophysics.* A.K. Peters.

Czikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. *Praha: Lidové Noviny.*

da Fabriano, G. (1423). *Adoration of the magi ,http://arthistoryblogger.blogspot.com/2011/11/continuous-narrative-in-art.html.*

Damala, A., Cubaud, P., Bationo, A., Houlier, P., & Marchal, I. (2008). Bridging the gap between the digital and the physical: design and evaluation of a mobile augmented reality guide for the museum visit. In *Proceedings of the 3rd international conference on digital interactive media in entertainment and arts* (pp. 120–127). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1413634.1413660` doi: http://doi.acm.org/10.1145/1413634.1413660

dAntonio, B. (1485). *The story of joseph. http://www.getty.edu/education/teachers/classroom_resources/curricula/stories_in_art/downloads/sia_story_joseph.pdf.*

DeCarlo, D., & Santella, A. (2002). Stylization and abstraction of photographs. In *Siggraph '02: Proceedings of the 29th annual conference on computer graphics and interactive techniques* (pp. 769–776). New York, NY, USA: ACM Press. doi: http://doi.acm.org/10.1145/566570.566650

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

Desser, T. S. (2007). Simulation-based training: the next revolution in radiology education? *J Am Coll Radiol*, *4*(11), 816–824.

Desurvire, H., Caplan, M., & Toth, J. A. (2004). Using heuristics to evaluate the playability of games. In *Chi'04 extended abstracts on human factors in computing systems* (pp. 1509–1512).

Dhawan, A., Chitre, Y., & Kaiser-Bonasso, C. (1996, jun). Analysis of mammographic microcalcifications using gray-level image structure features. *Medical Imaging, IEEE Transactions on*, *15*(3), 246 -259. doi: 10.1109/42.500063

di Buoninsegna, D. (1311a). *Christ taken prisoner http://www.wga.hu/frames-e.html?/html/d/duccio/.*

di Buoninsegna, D. (1311b). *Maesta alterpiece, http://www.casasantapia.com/art/duccio/maestabackpanels.htm.*

Dodge, R. (1900). Visual perception during eye movement. *Psychological Review*, *7*, 454-465.

Dodge, R., & Cline, T. (1901). The angle velocity of eye movements. *Psychological Review*, *8*, 145-157.

Dorr, M., Martinetz, T., Gegenfurtner, K., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of vision*, *10*(10).

Doshi, A., & Trivedi, M. (2010). Attention estimation by simultaneous observation of viewer and view. In *Computer vision and pattern recognition workshops (cvprw), 2010 ieee computer society conference on* (pp. 21–27).

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, *92*(1), 53–78.

Duchowski, A. (2007). *Eye tracking methodology: Theory and practice* (Vol. 373). Springer.

Duchowski, A. T. (2002, November). A breadth-first survey of eye-tracking applications. *Behav Res Methods Instrum Comput*, *34*(4), 455–470.

Duchowski, A. T., & Çöltekin, A. (2007). Foveated gaze-contingent displays for peripheral lod management, 3d visualization, and stereo imaging. *ACM Trans. Multimedia Comput. Commun. Appl.*, *3*(4), 1–18. doi: http://doi.acm.org/10.1145/1314303.1314309

Dugas, M., Trumm, C., Stäbler, A., Pander, E., Hundt, W., Scheidler, J., ... Reiser, M. (2001). Case-oriented computer-based-training in radiology: concept, implementation and evaluation. *BMC Med Educ*, *1*, 5.

Eckstein, M., Peterson, M., Pham, B., & Droll, J. (2009). Statistical decision theory to relate neurons to behavior in the study of covert visual attention. *Vision research*, *49*(10), 1097–1128.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, *9*(2), 111–118.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and bayesian priors. *Psychological Science*, *17*(11), 973–980.

Eckstein, M. P., Shimozaki, S. S., & Abbey, C. K. (2002). The footprints of visual attention in the posner cueing paradigm revealed by classification images. *Journal of Vision*, *2*(1).

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6-7), 945–978.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2).

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14).

El-Baz, A., Farag, A., Gimel?farb, G., Falk, R., El-Ghar, M. A., & Eldiasty, T. (2006). A framework for automatic segmentation of lung nodules from low dose chest ct scans. *Pattern Recognition, International Conference on*, *3*, 611-614. doi: http://doi.ieeecomputersociety.org/10.1109/ICPR.2006.66

El-Nasr, M. S., & Yan, S. (2006). Visual attention in 3d video games. In *Proceedings of the 2006 acm sigchi international conference on advances in computer entertainment technology* (p. 22).

Fawcett, T. (2004). *Roc graphs: Notes and practical considerations for researchers* (Tech. Rep.). Hewlett Packard.

Feil, J., & Scattergood, M. (2005). *Beginning game level design*. Course Technology PTR.

Feiner, S., MacIntyre, B., Hollerer, T., & Webster, A. (1997). A touring machine: Prototyping 3d mobile augmented reality systems for exploring the urban environment. In *Proceedings of the 1st ieee international symposium on wearable computers* (pp. 74–). Washington, DC, USA: IEEE Computer Society.

Forssén, P., Meger, D., Lai, K., Helmer, S., Little, J., & Lowe, D. (2008). Informed visual search: Combining attention and object recognition. In *Robotics and automation, 2008. icra 2008. ieee international conference on* (pp. 935–942).

Foulsham, T. (2012). 'eyes closed' and 'eyes open' expectations guide fixations in real-world search. In *Proc. cogsci 2012* (p. 1-6).

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception - London*, *36*(8), 1123.

Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (Vol. 3899). Springer-Verlag New York Inc.

Frintrop, S., Backer, G., & Rome, E. (2005). Goal-directed search with a top-down modulated computational attention system. *Pattern Recognition*, 117–124.

Frintrop, S., Rome, E., & Christensen, H. (2010a). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, *7*(1), 6.

Frintrop, S., Rome, E., & Christensen, H. I. (2010b). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, *7*(1), 6.

Gajewski, D. A., Pearson, A., Mack, M. L., Bartlett III, F. N., & Henderson, J. M. (2005). Lecture notes in computer science. In (Vol. 3368/2005, p. 83-99). Springer Berlin - Heidelberg.

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*(6), 989–1005.

Gao, D., & Vasconcelos, N. (2004). Discriminant saliency for visual recognition from cluttered scenes. *Advances in neural information processing systems*, *17*(481-488), 1.

Garcia-Diaz, A., Fdez-Vidal, X., Pardo, X., & Dosil, R. (2009). Decorrelation and distinctiveness provide with human-like saliency. In *Advanced concepts for intelligent vision systems* (pp. 343–354).

Gay, S. B., Sobel, A. H., Young, L. Q., & Dwyer, S. J. (1997). Processes involved in reading imaging studies: workflow analysis and implications for workstation develoment. *J Digital Imaging*, *10*, 40-45.

GD, T., R, V.-V., Jr, C. D., & Floyd CE, J. (2003). Computer-assisted detection of mammographic masses: a template matching scheme based on mutual information. *Medical Physics*, *30*, 2123–2130.

Goferman, S., Zelnik-Manor, L., & Tal, A. (2010). Context-aware saliency detection. In *Computer vision and pattern recognition (cvpr), 2010 ieee conference on* (pp. 2376–2383).

Gonzalez, R. C., & Woods, R. E. (2006). *Digital image processing (3rd edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Google Inc. (2014). *Google™ Image Search.* (http://images.google.com/)

Gottlieb, J., & Balan, P. (2010). Attention as a decision in information space. *Trends in cognitive sciences*, *14*(6), 240–248.

Grant, E., & Spivey, M. J. (2003). Eye movements and problem solving: guiding attention guides thought. *Psychological Science*, *14*(5), 462-466.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Greene, M., Liu, T., & Wolfe, J. (2012). Reconsidering yarbus: A failure to predict observers task from eye movement patterns. *Vision Research*.

Grillon, H., & Thalmann, D. (2009). Simulating gaze attention behaviors for crowds. *Computer Animation and Virtual Worlds*, *20*(2-3), 111–119.

Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on* (pp. 1–8).

Guo, C., & Zhang, L. (2010). A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans. on Image Processing.*, *19*.

Gwilt, I. (2009). Augmented reality and mobile art. In B. Furht (Ed.), *Handbook of multimedia for digital entertainment and arts* (p. 593-599). Springer US.

Hallowell, B., & Lansing, C. R. (2001). *Tracking eye movements to study cognition and communication.*

Hansen, L., Karadogan, S., & Marchegiani, L. (2011). What to measure next to improve decision making? on top-down task driven feature saliency. In *Computational intelligence, cognitive algorithms, mind, and brain (ccmb), 2011 ieee symposium on* (pp. 1–7).

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in neural information processing systems*, *19*, 545.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1).

Henderson, J., Brockmole, J., Castelhano, M., & Mack, M. (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., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, *50*(1), 243–271.

Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (p. 269-293). Oxford: Elsevier.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, *50*(1), 243–271.

Henderson, J. M., Weeks Jr, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, *25*(1), 210.

Hillaire, S., Lécuyer, A., Regia-Corte, T., Cozot, R., Royan, J., & Breton, G. (2010). A real-time visual attention model for predicting gaze point during first-person exploration of virtual environments. In *Proceedings of the 17th acm symposium on virtual reality software and technology* (pp. 191–198).

Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Ieee conference on computer vision and pattern recognition (cvpr).*

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments. *Advances in Neural Information Processing Systems (NIPS).*, 681–688.

Huey, E. (1968). *The psychology and pedagogy of reading.* Cambridge, MA: MIT Press. ((Originally published 1908))

Hurvich, L. M., & Jameson, D. (1957). An opponent-process theory of color vision. *Psychological Review*, *64*, 384-404.

Hwang, A., Wang, H., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision research.*

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision research*, *51*(10), 1192–1205.

Itti, L. (2006, Aug-Dec). Quantitative modeling of perceptual salience at human eye position [mod ; bu ; eye]. *Visual Cognition*, *14*(4-8), 959-984.

Itti, L., & Baldi, P. F. (2005, Jun). A principled approach to detecting surprising events in video [bu ; cv ; eye ; su]. In *Proc. ieee conference on computer vision and pattern recognition (cvpr)* (p. 631-637). San Siego, CA.

Itti, L., Dhavale, N., & Pighin, F. (2003, Aug). Realistic avatar eye and head animation using a neurobiological model of visual attention [bu;mod;cv]. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.), *Proc. spie 48th annual international symposium on optical science and technology* (Vol. 5200, p. 64-78). Bellingham, WA: SPIE Press.

Itti, L., Gold, C., & Koch, C. (2001, Sep). Visual attention and target detection in cluttered natural scenes [bu ; mod ; cv]. *Optical Engineering*, *40*(9), 1784-1793.

Itti, L., & Koch, C. (2000a, May). A saliency-based search mechanism for overt and covert shifts of visual attention [bu;mod;cv]. *Vision Research*, *40*(10-12), 1489-1506.

Itti, L., & Koch, C. (2000b, May). A saliency-based search mechanism for overt and covert shifts of visual attention [bu;mod;cv]. *Vision Research*, *40*(10-12), 1489-1506.

Itti, L., & Koch, C. (2001a). Computational modelling of visual attention. *Nature reviews neuroscience*, *2*(3), 194–203.

Itti, L., & Koch, C. (2001b, Mar). Computational modelling of visual attention [mod;bu;td;rev]. *Nature Reviews Neuroscience*, *2*(3), 194-203.

Itti, L., & Koch, C. (2001c, Mar). Computational modelling of visual attention [mod;bu;td;rev]. *Nature Reviews Neuroscience*, *2*(3), 194-203.

Itti, L., & Koch, C. (2001d, Jan). Feature combination strategies for saliency-based visual attention systems [bu;mod;cv]. *Journal of Electronic Imaging*, *10*(1), 161-169.

Itti, L., Koch, C., & Niebur, E. (1998, Nov). A model of saliency-based visual attention for rapid scene analysis [mod;bu;cv]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254-1259.

Itti, L., Koch, C., Niebur, E., et al. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(11), 1254–1259.

Jacob, R. J. K., & Karn, K. S. (2003). *The mind's eye: Cognitive and applied aspects of eye movement research.* Amsterdam: Elsevier Science.

Jensen, H. W., Marschner, S. R., Levoy, M., & Hanrahan, P. (2001). A practical model for subsurface light transport. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 511–518).

Jodogne, S., & Piater, J. (2007). Closed-loop learning of visual control policies. *Journal of Artificial Intelligence Research*, *28*(1), 349–391.

Johnson, D., & Wiles, J. (2003). Effective affective user interface design in games. *Ergonomics*, *46*(13-14), 1332–1345.

Ju, E., & Wagner, C. (1997). Personal computer adventure games: their structure, principles, and applicability for training. *ACM SIGMIS Database*, *28*(2), 78–92.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009a). Learning to predict where humans look. In *Computer vision, 2009 ieee 12th international conference on* (pp. 2106–2113).

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009b). Learning to predict where humans look. In *Computer vision, 2009 ieee 12th international conference on* (pp. 2106–2113).

Kaneko, T., Takahei, T., Inami, M., Kawakami, N., Yanagida, Y., Maeda, T., & Tachi, S. (2001). Detailed shape representation with parallax mapping. In *Proceedings of icat* (Vol. 2001, pp. 205–208).

Kastner, S., Pinsk, M., De Weerd, P., Desimone, R., & Ungerleider, L. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, *22*(4), 751–761.

Kienzle, W., Wichmann, F., Schölkopf, B., & Franz, M. (2007). A nonparametric approach to bottom-up visual saliency. In *Proc. nips* (p. 1-8).

Kim, Y., & Varshney, A. (2006). Saliency-guided enhancement for volume visualization. *IEEE Trans. on Visualization and Computer Graphics (IEEE Visualization 2006)*, *12, No. 5*, 925–932.

Kimura, A., Pang, D., Takeuchi, T., Yamato, J., & Kashino, K. (2008). Dynamic markov random fields for stochastic modeling of visual attention. In *Pattern recognition, 2008. icpr 2008. 19th international conference on* (pp. 1–5).

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology.*, *4*(4), 219–227.

Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115–141). Springer.

Koffka, K. (1935). *Principles of gestalt psychology.* Harcourt, Brace and World, New York, NY.

Kootstra, G., Nederveen, A., & De Boer, B. (2008). Paying attention to symmetry. In *Proceedings of the british machine vision conference (bmvc2008)* (pp. 1115–1125).

Kosara, R., Miksch, S., & Hauser, H. (2001, 22–23). Semantic depth of field. In *IEEE symposium on information visualization 2001 (infovis 2001).* San Diego, CA, USA. Retrieved from `citeseer.ist.psu.edu/article/kosara01semantic.html`

Koulieris, G. A., Drettakis, G., Cunningham, D., & Mania, K. (2014a). An automated high level saliency predictor for smart game balancing. *ACM Transactions on Applied Perception (TAP).*

Koulieris, G. A., Drettakis, G., Cunningham, D., & Mania, K. (2014b). C-lod: Context-aware material level-of-detail for mobile graphics. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering 2014, Lyon, France).*

Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Acad Radiol*, *3*(2), 137–144.

Krupinski, E. A. (2000). The importance of perception research in medical imaging. *Radiat Med*, *18*(6), 329-34.

Kundel, H. L., Nodine, C. F., Krupinski, E. A., & Mello-Thoms, C. (2008). Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Academic Radiology*, *15*(7), 881 - 886. Retrieved from `http://www.sciencedirect.com/` doi: DOI: 10.1016/j.acra.2008.01.023

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

Land, M., & McLeod, P. (2000). From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, *3*(12), 1340–1345.

Lee, C. H., Varshney, A., & Jacobs, D. (2005). Mesh saliency. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2005)*, *24, No. 3*, 659 – 666.

Lee, S., Kim, G. J., & Choi, S. (2009). Real-time tracking of visually attended objects in virtual environments and its application to lod. *Visualization and Computer Graphics, IEEE Transactions on*, *15*(1), 6–19.

Lee, S., Oh, J., Park, J., Kwon, J., Kim, M., & Yoo, H. (2011). A 345 mw heterogeneous many-core processor with an intelligent inference engine for robust object recognition. *Solid-State Circuits, IEEE Journal of*, *46*(1), 42–51.

Le Meur, O., Le Callet, P., Barba, D., & Thoreau, D. (2006). A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *28*(5), 802–817.

Levenshtein, V. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, *1*, 8-17.

Levenshtein, V. I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals* (Vol. 10; Tech. Rep. No. 8). Moscow State University.

Levine, J. (1981). *An eye-controlled computer* (Research Report No. RC-8857). Yorktown Heights, N.Y.: IBM Thomas J. Watson Research Center.

Li, F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596.

Li, H., Kallergi, M., Clarke, L., Jain, V., & Clark, R. (1995, sep). Markov random field for tumor detection in digital mammography. *Medical Imaging, IEEE Transactions on*, *14*(3), 565 -576. doi: 10.1109/42.414622

Li, J., Tian, Y., Huang, T., & Gao, W. (2011). Multi-task rank learning for visual saliency estimation. *Circuits and Systems for Video Technology, IEEE Transactions on*(99), 1–1.

Li, L., & Fei-Fei, L. (2010). Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, *88*(2), 147–168.

Li, Y., Zhou, Y., Yan, J., Niu, Z., & Yang, J. (2010). Visual saliency based on conditional entropy. *Computer Vision–ACCV 2009*, 246–257.

Lindholm, E., Kilgard, M. J., & Moreton, H. (2001). A user-programmable vertex engine. In *Proceedings of the 28th annual conference on computer graphics and interactive techniques* (pp. 149–158).

Litchfield, D., & Ball, L. J. (2011, Apr). Using another's gaze as an explicit aid to insight problem solving. *Q J Exp Psychol (Hove)*, *64*, 649–656.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2010, Sep). Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection. *J Exp Psychol Appl*, *16*, 251–262.

Liu, T., Sun, J., Zheng, N., Tang, X., & Shum, H. (2007). Learning to detect a salient object. *IEEE Conference on Computer Vision and Pattern Recognition.*.

Livingstone, M. (2002). *Vision and art: The biology of seeing*. Harry N. Abrams, Inc.

Longhurst, P., Debattista, K., & Chalmers, A. (2006). A gpu based saliency map for high-fidelity selective rendering. In *Proceedings of the 4th international conference on computer graphics, virtual reality, visualisation and interaction in africa* (pp. 21–29).

Loschky, L. C., & McConkie, G. W. (2000). User performance with gaze contingent multiresolutional displays. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 97–103).

Luebke, D., & Erikson, C. (1997). View-dependent simplification of arbitrary polygonal environments. *Computer Graphics*, *31*(Annual Conference Series), 199–208. Retrieved from `citeseer.ist.psu.edu/21067.html`

Luebke, D., Watson, B., Cohen, J. D., Reddy, M., & Varshney, A. (2002). *Level of detail for 3d graphics*. New York, NY, USA: Elsevier Science Inc.

Luebke, D. P. (2003). *Level of detail for 3d graphics*. Morgan Kaufmann.

Ma, X., & Deng, Z. (2009). Natural eye motion synthesis by modeling gaze-head coupling. In *Virtual reality conference, 2009. vr 2009. ieee* (pp. 143–150).

Ma, Y., Hua, X., Lu, L., & Zhang, H. (2005). A generic framework of user attention model and its application in video summarization. *Multimedia, IEEE Transactions on*, *7*(5), 907–919.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. MIT: MIT Press.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*, 547-552.

Mahadevan, V., & Vasconcelos, N. (2010). Spatiotemporal saliency in dynamic scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*(1), 171–177.

MANCAS, M. (2007). *Computational attention towards attentive computers*. Presses univ. de Louvain.

Mancas, M., Gosselin, B., & Macq, B. (2004). Automatic fast detection of tumor suspect areas on ct scan. *Visualization Conference, IEEE*, *0*, 33p. doi: http://doi.ieeecomputersociety.org/10.1109/VISUAL.2004.9

Mania, K., Robinson, A., & Brandt, K. R. (2005). The effect of memory schemas on object recognition in virtual environments. *Presence: Teleoperators and Virtual Environments*, *14*(5), 606–615.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*, 165-188.

Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., & Guérin-Dugué, A. (2009). Modelling spatio-temporal saliency to predict gaze direction for short videos. *International Journal of Computer Vision*, *82*(3), 231–243.

Markou, M., & Singh, S. (2003). Novelty detection: a review part 2, neural network based approaches. *Signal processing*, *83*(12), 2499–2521.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*.

Martinez-Trujillo, J. (2011). Searching for the neural mechanisms of feature-based attention in the primate brain. *Neuron*, *70*(6), 1025–1028.

Masaccio, A. (1421). *The tribute money.* Retrieved from `http://en.wikipedia.org/wiki/The_Tribute_Money`

McCallum, A. (1996). *Reinforcement learning with selective perception and hidden state* (Unpublished doctoral dissertation). University of Rochester.

McNamara, A. (2011). Enhancing art history education through mobile augmented reality. In *Proceedings of the 11th acm siggraph international conference on virtual-reality continuum and its applications in industry* (p. 1). New York, NY, USA: ACM.

McNamara, A., Bailey, R., & Grimm, C. (2008a). Improving search task performance using subtle gaze direction. In *Proceedings of the 5th symposium on applied perception in graphics and visualization* (pp. 51–56). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1394281.1394289` doi: http://doi.acm.org/10.1145/1394281.1394289

McNamara, A., Bailey, R., & Grimm, C. (2008b). Improving search task performance using subtle gaze direction. In *Proceedings of the 5th symposium on applied perception in graphics and visualization* (pp. 51–56). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1394281.1394289` doi: http://doi.acm.org/10.1145/1394281.1394289

McNamara, A., Booth, T., Sridharan, S., Caffey, S., Grimm, C., & Bailey, R. (2012). Directing gaze in narrative art. In *Proceedings of the acm symposium on applied perception* (pp. 63–70). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2338676.2338689` doi: 10.1145/2338676.2338689

Mello-Thoms, C., Britton, C., Abrams, G., Hakim, C., Shah, R., Hardesty, L., ... Gur, D. (2006). Head-mounted versus remote eye tracking of radiologists searching for breast cancer: a comparison. *Acad Radiol*, *13*(2), 203–209.

Miau, F., Papageorgiou, C., & Itti, L. (2001, Nov). Neuromorphic algorithms for computer vision and attention [bu;mod;cv]. In B. Bosacchi, D. B. Fogel, & J. C. Bezdek (Eds.), *Proc. spie 46 annual international symposium on optical science and technology* (Vol. 4479, p. 12-23). Bellingham, WA: SPIE Press.

Milanese, R., Wechsler, H., Gill, S., Bost, J., & Pun, T. (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer vision and pattern recognition, 1994. proceedings cvpr'94., 1994 ieee computer society conference on* (pp. 781–785).

Mitchell, G. E. (2004). Taking control over depth of field: Using the lens blur filter in adobe photoshop cs [Computer software manual].

Moehring, M., Gloystein, A., & Doerner, R. (2009). Issues with virtual space perception within reaching distance: Mitigating adverse effects on applications using hmds in the automotive industry. In *Virtual reality conference, 2009. vr 2009. ieee* (pp. 223–226).

Mourant, R., & Rockwell, T. (1970). Mapping eye-movement patterns to the visual scene in driving: An exploratory study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *12*(1), 81–87.

Mourkoussis, N., Rivera, F. M., Troscianko, T., Dixon, T., Hawkes, R., & Mania, K. (2010). Quantifying fidelity for virtual environment simulations employing memory schema assumptions. *ACM Transactions on Applied Perception (TAP)*, *8*(1), 2.

Muniyandi, M., Cotin, S., Srinivasan, M., & Dawson, S. (2003). Real-time pc based x-ray simulation for interventional radiology training. *Stud Health Technol Inform*, *94*, 233-9.

Murphy, H., & Duchowski, A. T. (2007). Hybrid image-/model-based gaze-contingent rendering. In *Apgv '07: Proceedings of the 4th symposium on applied perception in graphics and visualization* (pp. 107–114). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/1272582.1272604

Murray, N., Vanrell, M., Otazu, X., & Parraga, C. (2011). Saliency estimation using a non-parametric low-level vision model. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on* (pp. 433–440).

Nakayama, K., Silverman, G. H., et al. (1986). Serial and parallel processing of visual feature conjunctions. *Nature*, *320*(6059), 264–265.

Navalpakkam, V., Arbib, M. A., & Itti, L. (2005, Jan). Attention and scene understanding [bu;td;psy;mod;sc]. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (p. 197-203). San Diego, CA: Elsevier.

Navalpakkam, V., & Itti, L. (2005, Jan). Modeling the influence of task on attention [bu ; td ; mod ; sc]. *Vision Research*, *45*(2), 205-231.

Navalpakkam, V., & Itti, L. (2006, Jun). An integrated model of top-down and bottom-up attention for optimal object detection [bu ; cv ; td]. In *Proc. ieee conference on computer vision and pattern recognition (cvpr)* (p. 2049-2056). New York, NY.

Navalpakkam, V., & Itti, L. (2007, Feb). Search goal tunes visual features optimally [bu;td;mod;psy]. *Neuron*, *53*(4), 605-617. (Also see commentary / preview entitled "Paying Attention to Neurons with Discriminating Taste" by A. Pouget and D. Bavelier, Neuron 2007;53(4):473-475.)

Navarro, G. (2001, March). A guided tour to approximate string matching. *ACM Comput. Surv.*, *33*(1), 31–88. Retrieved from `http://doi.acm.org/10.1145/375360.375365` doi: 10.1145/375360.375365

NMC. (2012, July). *Horizon report, museum edition 2011.* Retrieved from `http://www.nmc.org/news/`

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, *171*(968), 308-11.

O'Craven, K. M., Downing, P. E., & Kanwisher, N. (1999). fmri evidence for objects as the units of attentional selection. *Nature*, *401*(6753), 584–587.

Ogden, T. E., & Miller, R. F. (1966a). Studies of the optic nerve of the rhesus monkey: Nerve fiber spectrum and physiological properties. *Vision Research*, *6*, 485-506.

Ogden, T. E., & Miller, R. F. (1966b). Studies of the optic nerve of the rhesus monkey: Nerve fiber spectrum and physiological properties. *Vision Research*, *6*(1), 485–506.

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detection. In *Image processing, 2003. icip 2003. proceedings. 2003 international conference on* (Vol. 1, pp. I–253).

Orabona, F., Metta, G., & Sandini, G. (2005). Object-based visual attention: a model for a behaving robot. In *Computer vision and pattern recognition-workshops, 2005. cvpr workshops. ieee computer society conference on* (pp. 89–89).

Osterberg, G. (1935). *Topography of the layer of rods and cones in the human retina.* Acta Ophthal. (suppl. 6, 11-97)

O'Sullivan, C., Dingliana, J., & Howlett, S. (2003). *Eye-movements and interactive graphics.* Retrieved from `citeseer.ist.psu.edu/570078.html`

Ouerhani, N., & Hügli, H. (2003). Real-time visual attention on a massively parallel simd architecture. *Real-Time Imaging*, *9*(3), 189–196.

Owens, J. (2005). Streaming architectures and technology trends. In *Acm siggraph 2005 courses* (p. 9).

Oyekoya, O., Steptoe, W., & Steed, A. (2009). A saliency-based method of simulating visual attention in virtual scenes. In *Proceedings of the 16th acm symposium on virtual reality software and technology* (pp. 199–206).

P., C. T., & W., M. D. (1992, January). Augmented reality: an application of heads- up display technology to manual manufacturing processes. *In Proceedings of the Twenty-Fifth Hawaii International Conference on Systems Science*, *2*(3), 659–669.

Pagulayan, R. J., Keeker, K., Wixon, D., Romero, R. L., & Fuller, T. (2003). User-centered design in games. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, 883–906.

Paletta, L., Rome, E., & Buxton, H. (2005). Attention architectures for machine vision and mobile robots. *Neurobiology of attention*, 642–648.

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* The MIT press.

Pang, D., Kimura, A., Takeuchi, T., Yamato, J., & Kashino, K. (2008). A stochastic model of selective visual attention with a dynamic bayesian network. In *Multimedia and expo, 2008 ieee international conference on* (pp. 1073–1076).

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

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

Peters, R. J., & Itti, L. (2007, Jun). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention [bu ; cv ; td ; eye ; mod]. In *Proc. ieee conference on computer vision and pattern recognition (cvpr).* Minneapolis, MN.

Peters, R. J., & Itti, L. (2008, Jun). Congruence between model and human attention reveals unique signatures of critical visual events [bu ; cv ; td ; eye ; mod]. In *Advances in neural information processing systems, vol. 20 (nips*2007)* (p. 1145-1152). Cambridge, MA: MIT Press.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005, Aug). Components of bottom-up gaze allocation in natural images [bu;td;eye;mod]. *Vision Research*, *45*(8), 2397-2416.

Posner, M. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and performance X: Control of language processes*, *32*, 531–556.

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(9), 970–982.

Qvarfordt, P., Biehl, J. T., Golovchinsky, G., & Dunningan, T. (2010). Understanding the benefits of gaze enhanced visual search. In *Proceedings of the 2010 symposium on eye-tracking research & applications* (pp. 283–290). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1743666.1743733` doi: http://doi.acm.org/10.1145/1743666.1743733

Rajashekar, U., van der Linde, I., Bovik, A., & Cormack, L. (2008). Gaffe: A gaze-attentive fixation finding engine. *Image Processing, IEEE Transactions on*, *17*(4), 564–573.

Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T. (2010). An eye fixation database for saliency detection in images. *Computer Vision–ECCV 2010*, 30–43.

Ramström, O., & Christensen, H. (2002). Visual attention using game theory. In *Biologically motivated computer vision* (pp. 462–471).

Rao, R., Zelinsky, G., Hayhoe, M., & Ballard, D. (1996). Modeling saccadic targeting in visual search. *Advances in neural information processing systems*, 830–836.

Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M., & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision research*, *42*(11), 1447–1463.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, *7*, 65-81.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, *62*(8), 1457–1506.

Relan, J., Sermesant, M., Pop, M., Delingette, H., Sorine, M., Wright, G., & Ayache, N. (2009). Volumetric prediction of cardiac electrophysiology using a heart model personalised to surface data. In *Miccai workshop on cardiovascular interventional imaging and biophysical modelling* (pp. 19–27). London, UK. Retrieved from `http://hal.inria.fr/inria-00418471/en/`

Renninger, L., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in neural information processing systems*, *17*, 1121–1128.

Rensink, R. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*(1-3), 17–42.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*, *7*(1-3), 17–42.

Reynolds, J., & Desimone, R. (1999). The role of neural mechanisms review of attention in solving the binding problem. *Neuron*, *24*, 19–29.

Rezazadegan Tavakoli, H., Rahtu, E., & Heikkilä, J. (2011). Fast and efficient saliency detection using sparse sampling and kernel density estimation. *Image Analysis*, 666–675.

Robertson, L., et al. (2003). Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, *4*(2), 93–102.

Rosin, P. (2009). A simple method for detecting salient regions. *Pattern Recognition*, *42*(11), 2363–2371.

Rothenstein, A., & Tsotsos, J. (2008). Attention links sensing to recognition. *Image and Vision Computing*, *26*(1), 114–126.

Rybak, I. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., & Shevtsova, N. A. (1998). A model of attention-guided visual perception and recognition. *Vision Res*, *38*(15-16), 2387-2400.

Saalmann, Y., Pinsk, M., Wang, L., Li, X., & Kastner, S. (2012). The pulvinar regulates information transmission between cortical areas based on attention demands. *Science*, *337*(6095), 753–756.

Saenz, M., Buracas, G. T., Boynton, G. M., et al. (2002). Global effects of feature-based attention in human visual cortex. *Nature neuroscience*, *5*(7), 631–632.

Salah, A., Alpaydin, E., & Akarun, L. (2002). A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(3), 420–425.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on eye tracking research & applications* (pp. 71–78).

Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., & Zetzsche, C. (2001). Knowledge-based scene analysis with saccadic eye-movements. *J. Electron. Imaging*, *10*(1), 152–160.

Schmidt, J., & Zelinsky, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology*, *62*(10), 1904–1914.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *Proceedings IEEE ICCST*, *38*, 258–264.

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements. In *Apgv '07: Proceedings of the 4th symposium on applied perception in graphics and visualization* (pp. 123–130). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/1272582.1272606

Secord, A., Lu, J., Finkelstein, A., Singh, M., & Nealen, A. (2011). Perceptual models of viewpoint preference. *ACM Transactions on Graphics (TOG)*, *30*(5), 109.

Seo, H., & Milanfar, P. (2009). Static and space-time visual saliency detection by self-resemblance. *Journal of Vision.*, *9*(12), 1–27.

Serences, J., & Saproo, S. (2010). Population response profiles in early visual cortex are biased in favor of more valuable stimuli. *Journal of neurophysiology*, *104*(1), 76–87.

Sharples, M., Jeffery, N. P., du Boulay, B., Teather, B. A., Teather, D., & du Boulay, G. H. (2000). Structured computer-based training in the interpretation of neuroradiological images. *International Journal of Medical Informatics*, *60*(3), 263 - 280. doi: DOI: 10.1016/S1386-5056(00)00101-5

Shic, F., & Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision*, *73*(2), 159–177.

Shipley, T. F., & Kellman, P. J. (2001). *From fragments to objects: Segmentation and grouping in vision* (Vol. 130). Access Online via Elsevier.

Siagian, C., & Itti, L. (2007, Feb). Rapid biologically-inspired scene classification using features shared with visual attention [mod;bu;sc]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(2), 300-312.

Simons, D., & Chabris, C. (1995). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *British Journal of Developmental Psychology*, *13*, 113–142. Retrieved from `citeseer.ist.psu.edu/simons99gorillas.html`

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in cognitive sciences*, *1*(7), 261–267.

Slater, M., Spanlang, B., & Corominas, D. (2010). Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Transactions on Graphics (TOG)*, *29*(4), 92.

Sodhi, M., Reimer, B., Cohen, J., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002a). On-road driver eye movement tracking using head-mounted devices. In *Proceedings of the 2002 symposium on eye tracking research & applications* (pp. 61–68).

Sodhi, M., Reimer, B., Cohen, J. L., Vastenburg, E., Kaars, R., & Kirschenbaum, S. (2002b). On-road driver eye movement tracking using head-mounted devices. In *Etra '02: Proceedings of the 2002 symposium on eye tracking research & applications* (pp. 61–68). New York, NY, USA: ACM. doi: http://doi.acm.org/10.1145/507072.507086

Sowden, P. T., Davies, I. R., & Roling, P. (2000, Feb). Perceptual learning of the detection of features in x-ray images: a functional role for improvements in adults' visual sensitivity? *J Exp Psychol Hum Percept Perform*, *26*(1), 379-90.

Spillman, L. (1990). *Visual perception: The neurophysiological foundations.* Academic Press, San Diego.

Spillmann, L. (1990). *Visual perception: The neurophysiological foundations.* San Diego: Academic Press.

Sprague, N., & Ballard, D. (2003). Eye movements for reward maximization. *Advances in neural information processing systems*, *16*.

Sridharan, S., Bailey, R., McNamara, A., & Grimm, C. (2011). Subtle gaze manipulation for improved mammography training. In *Proceedings of the acm siggraph symposium on applied perception in graphics and visualization* (pp. 112–112). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2077451.2077475` doi: 10.1145/2077451.2077475

Srinivasan, R., Boast, R., Furner, J., & Becvar, K. M. (2009, July). Digital museums and diverse cultural knowledges: Moving past the traditional catalog. *The Information Society*, *25*, 265–278. Retrieved from `http://portal.acm.org/citation.cfm?id=1568347.1568352` doi: 10.1080/01972240903028714

Steinmetz, P. N., Roy, A., Fitzgerald, P., Hsiao, S., Johnson, K., & Niebur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature*, *404*(6774), 187–190.

Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., ... Savage, J. (1994). The mammographic image analysis society digital mammogram database. In *International congress series* (Vol. 1069, p. 375-378).

Summerfield, J., Rao, A., Garside, N., & Nobre, A. (2011). Biasing perception by spatial long-term memory. *The Journal of Neuroscience*, *31*(42), 14952–14960.

Sun, X., Yao, H., & Ji, R. (2012). What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *Proc ieee-cvpr* (p. 1552-1559).

Sun, Y., & Fisher, R. (2003). Object-based visual attention for computer vision. *Artificial Intelligence*, *146*(1), 77–123.

Sun, Y., Fisher, R., Wang, F., & Gomes, H. (2008). A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding*, *112*(2), 126–142.

Sundstedt, V., Bernhard, M., Stavrakis, E., Reinhard, E., & Wimmer, M. (2013). Visual attention and gaze behavior in games: An object-based approach. In *Game analytics* (pp. 543–583). Springer.

Sundstedt, V., Chalmers, A., Cater, K., & Debattista, K. (2004). Top-down visual attention for efficient rendering of task related scenes. In *Vmv* (pp. 209–216).

Sundstedt, V., Debattista, K., Longhurst, P., Chalmers, A., & Troscianko, T. (2005). Visual attention for efficient high-fidelity graphics. In *Proceedings of the 21st spring conference on computer graphics* (pp. 169–175).

Sundstedt, V., Stavrakis, E., Wimmer, M., & Reinhard, E. (2008). A psychophysical study of fixation behavior in a computer game. In *Proceedings of the 5th symposium on applied perception in graphics and visualization* (pp. 43–50).

Sweetser, P., & Wyeth, P. (2005). Gameflow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, *3*(3), 3–3.

Tatarchuk, N. (2006). Dynamic parallax occlusion mapping with approximate soft shadows. In *Proceedings of the 2006 symposium on interactive 3d graphics and games* (pp. 63–69).

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision research*, *45*(5), 643–659.

Tatler, B., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, *11*(5).

Tatler, B., & Vincent, B. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029–1054.

Theeuwes, J. (2010a). Top–down and bottom–up control of visual selection. *Acta psychologica*, *135*(2), 77–99.

Theeuwes, J. (2010b). Top–down and bottom–up control of visual selection. *Acta psychologica*, *135*(2), 77–99.

Theeuwes, J., & Godijn, R. (2002). Irrelevant singletons capture attention: Evidence from inhibition of return. *Perception & Psychophysics*, *64*(5), 764–770.

Thomas, L., & Lleras, A. (2007, August). Moving eyes and moving thought: on the spatial compatibility between eye movements and cognition. *Psychonomic bulletin and review*, *14*(4), 663-668.

Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *33*(11), 2131–2146.

Tolhurst, D. J., Movshon, J. A., & Dean, A. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision research*, *23*(8), 775–785.

Torralba, A. (2003). Modeling global scene factors in attention. *JOSA A*, *20*(7), 1407–1418.

Torralba, A., & Efros, A. (2011). Unbiased look at dataset bias. In *Ieee conference on computer vision and pattern recognition.*

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology.*, *12*, 97–136.

Treisman, A., & Souther, J. (1985). Search asymmetry: a diagnostic for preattentive processing of separable features. *J Exp Psychol Gen*, *114*(3), 285-310.

Treisman, A. M., & Gelade, G. (1980a). A feature-integration theory of attention. *Cognit Psychol*, *12*(1), 97-136.

Treisman, A. M., & Gelade, G. (1980b). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Treue, S., & Martinez Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, *399*(6736), 575-579.

Tseng, P., Carmi, R., Cameron, I., Munoz, D., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7).

Tsotsos, J., & Rothenstein, A. (2011). Computational models of visual attention. *Scholarpedia*, *6*(1), 6201.

Tu, Z., Zhou, X. S., Bogoni, L., Barbu, A., & Comaniciu, D. (2006a). Probabilistic 3d polyp detection in ct images: The role of sample alignment. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, *2*, 1544-1551. doi: http://doi.ieeecomputersociety.org/10.1109/CVPR.2006.228

Tu, Z., Zhou, X. S., Bogoni, L., Barbu, A., & Comaniciu, D. (2006b). Probabilistic 3d polyp detection in ct images: The role of sample alignment. In *Cvpr '06: Proceedings of the 2006 ieee computer society conference on computer vision and pattern recognition* (pp. 1544–1551). Washington, DC, USA: IEEE Computer Society. doi: http://dx.doi.org/10.1109/CVPR.2006.228

Underwood, G., & Foulsham, T. (2006, November). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Q J Exp Psychol (Colchester)*, *59*(11), 1931–1949. Retrieved from `http://dx.doi.org/10.1080/17470210500416342` doi: 10.1080/17470210500416342

Veas, E. E., Mendez, E., Feiner, S. K., & Schmalstieg, D. (2011). Directing attention and influencing memory with visual saliency modulation. In *Proceedings of the 2011 annual conference on human factors in computing systems* (pp. 1471–1480). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/1978942.1979158` doi: http://doi.acm.org/10.1145/1978942.1979158

Velli, S. T. (2007). *Le immagini e il tempo: Narrazione visiva, storia e allegoria tra cinque e seicento.* Italy: Edizioni Della Normale.

Verma, M., & McOwan, P. (2009). Generating customised experimental stimuli for visual search using genetic algorithms shows evidence for a continuum of search efficiency. *Vision research*, *49*(3), 374–382.

Vickerss, J. (2007). *Perception, cognition, and decision training: The quiet eye in action.* New York, NY: Human Kinetics.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer vision and pattern recognition, 2001. cvpr 2001. proceedings of the 2001 ieee computer society conference on* (Vol. 1, pp. I–511).

Võ, M., & Henderson, J. (2009). Does gravity matter? effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3).

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395–1407.

Wang, H., Freeman, J., Merriam, E., Hasson, U., & Heeger, D. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*, *12*(1).

Wang, R. F., & Spelke, E. S. (2002). Human spatial representation: insights from animals. *Trends in Cognitive Sciences*, *6*(9), 376 - 382. doi: DOI: 10.1016/S1364-6613(02)01961-7

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

Ward, G. (2000). *Rendering with radiance.* Morgan Kaufmann.

Witz, K. (1434). *The miraculous draught of the fishes, http://mydailyartdisplay.wordpress.com/2011/07/18/the-miraculous-draught-of-the-fishes-by-konrad-witz-2/.* http://mydailyartdisplay.wordpress.com/2011/07/18/the-miraculous-draught-of-the-fishes-by-konrad-witz-2/.

Wolfe, J. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202–238.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202–238.

Wyman, C. (2005). An approximate image-space approach for interactive refraction. *ACM Transactions on Graphics (TOG)*, *24*(3), 1050–1053.

Xu, R., & Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. Retrieved from `http://dx.doi.org/10.1109/TNN.2005.845141` doi: 10.1109/TNN.2005.845141

Xu, T., Kuhnlenz, K., & Buss, M. (2010). Autonomous behavior-based switched top-down and bottom-up visual attention for mobile robots. *Robotics, IEEE Transactions on*, *26*(5), 947–954.

Yahoo! Inc. (2014). *flickr™*. (http://www.flickr.com/)

Yarbus, A. (1967). *Eye movements and vision.* New York: Plenum Press.

Yarbus, A. L. (1967). *Eye movements and vision.* Plenum Press. (Translated from Russian by Basil Haigh. Original Russian edition published in Moscow in 1965)

Yarbus, A. L., Haigh, B., & Rigss, L. A. (1967). *Eye movements and vision* (Vol. 2) (No. 5.10). Plenum press New York.

Yu, Y., Mann, G., & Gosine, R. (2008). An object-based visual attention model for robots. In *Robotics and automation, 2008. icra 2008. ieee international conference on* (pp. 943–948).

Yu, Y., Mann, G., & Gosine, R. (2012). A goal-directed visual perception system using object-based top-down attention. *Autonomous Mental Development, IEEE Transactions on*, *4*(1), 87-103.

Zelinsky, G., Zhang, W., Yu, B., Chen, X., & Samaras, D. (2006). The role of top-down and bottom-up processes in guiding eye movements during visual search. *Advances in neural information processing systems*, *18*, 1569.

Zhai, Y., & Shah, M. (2006). Visual attention detection in video sequences using spatiotemporal cues. In *Proceedings of the 14th annual acm international conference on multimedia* (pp. 815–824).

Zhang, L., Tong, M., & Cottrell, G. (2009). Sunday: Saliency using natural statistics for dynamic analysis of scenes. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2944–2949).

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

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3).

Zhou, H., & Desimone, R. (2011). Feature-based attention in the frontal eye field and area v4 during visual search. *Neuron*, *70*(6), 1205–1217.

Ziegler, G., Tevs, A., Theobalt, C., & Seidel, H.-P. (2006). On-the-fly point clouds through histogram pyramids. In *Workshop on vision, modeling, and visualization (vmv 2006)* (pp. 137–144).

Zotos, A., Mania, K., & Mourkoussis, N. (2009). A schema-based selective rendering framework. In *Proceedings of the 6th symposium on applied perception in graphics and visualization* (pp. 85–92).