Scenes, saliency maps and scanpaths

Tom Foulsham

Table of contents

7. Applications and implications
    a. Perception of art
    b. Marketing and websites
    c. Computer vision
8. Conclusions and limitations
9. Suggested readings
10. Questions for discussion
11. Bibliography

2. Summary and Learning objectives

The aim of this chapter is to review some of the key research investigating how people look at pictures. In particular, my goal is to provide theoretical background for those that are new to the field, while also explaining some of the relevant methods and analyses.

I begin by introducing eye movements in the context of natural scene perception. As in other complex tasks, eye movements provide a measure of attention and information processing over time, and they tell us about how the foveated visual system determines what to prioritise. I then describe some of the many measures which have been derived to summarize where people look in complex images. These include global measures, analyses based on regions of interest and comparisons based on heat maps.

A particularly popular approach for trying to explain fixation locations is the saliency map approach, and the first half of the chapter is mostly devoted to this topic. A large number of papers and models are built on this approach, but it is also worth spending time on this topic because the methods involved have been used across a wide range of applications. The saliency map approach is based on the fact that the visual system has topographic maps of visual features, that contrast within these features seems to be represented and prioritized, and that a central representation can be used to control attention and eye movements. This approach, and the underlying principles, has led to an increase in the number of researchers using complex natural scenes as stimuli. It is therefore important that those new to the field are familiar with saliency maps, their usage, and their pitfalls. I describe the original implementation of this approach (Itti & Koch, 2000), which uses spatial filtering at different levels of

coarseness and combines them in an attempt to identify the regions which stand out from their background. Evaluating this model requires comparing fixation locations to model predictions. Several different experimental and comparison methods have been used, but most recent research shows that bottom-up guidance is rather limited in terms of predicting real eye movements.

The second part of the chapter is largely concerned with measuring eye movement scanpaths. Scanpaths are the sequential patterns of fixations and saccades made when looking at something for a period of time. They show regularities which may reflect top-down attention, and some have attempted to link these to memory and an individual's mental model of what they are looking at. While not all researchers will be testing hypotheses about scanpaths, an understanding of the underlying methods and theory will be of benefit to all. I describe the theories behind analyzing eye movements in this way, and various methods which have been used to represent and compare them. These methods allow one to quantify the similarity between two viewing patterns, and this similarity is linked to both the image and the observer.

The last part of the chapter describes some applications of eye movements in image viewing. The methods discussed can be applied to complex images, and therefore these experiments can tell us about perception in art and marketing, as well as about machine vision.

By the end of the chapter, readers should

- Understand why eye movements are useful for studying natural scene perception.
- Understand some of the measures used for quantifying fixations on images.

- Appreciate the theoretical and neural underpinnings of a saliency map approach.
- Understand the Itti and Koch (2000) model of bottom-up visual saliency.
- Be able to evaluate saliency map models using eye fixation data.
- Appreciate temporal aspects of eye movements in scene, including fixation duration and order
- Understand scanpaths and why they have been studied.
- Appreciate comparison methods which look at the scanpath sequence.
- Be familiar with some of the applications of eyetracking experiments in images.

3. Introduction

Eye movements are a fundamental part of natural human vision. This is particularly true when we consider complex natural scenes, which comprise more information than we can take in within a single glance. Due to our sampling of the visual field, which is dominated by high acuity at the fovea and decreased resolution everywhere else on the retina, we can only inspect our environment in detail by moving our eyes and bodies to select different regions of interest. This process of actively selecting information gives psychology and neuroscience a uniquely sensitive measure about how people perceive and understand images. However, it also creates difficulties for a visual brain which has to rapidly orient the eyes based on only peripheral information, and then combine the input from multiple fixations so that we can understand a scene and act accordingly.

This chapter describes some of the theory and methods used in the study of eye movements in complex stimuli. I will focus on two particular sets of theoretical questions within this topic, which are related to *saliency maps* and *scanpaths*. However, this chapter could easily be called "Looking at pictures", because the research and methods being discussed are those where we measure people looking at pictures and photographs of scenes. Broadly speaking, the research to be discussed tries to describe *where* people look and *in what order*. Explaining these two things is a complex problem, but it should be easy to see that doing so will involve both the visual appearance of items in the scene (e.g., how bright or colourful something is) and the knowledge or task of the observer. It is worth bearing in mind, from the outset, that these depictions of the world are convenient abstractions for

experimenters, but that they may not always reflect the way that we move our eyes in the real environment.

When investigating the viewing of static images, most researchers analyse saccades and fixations (for background on the properties of saccades, see chapters by Pierce et al. and Hutton, this volume). Although scene viewing may elicit other eye movement events, such as fixational eye movements and microsaccades (see Martinez-Conde & Alexander, this volume), saccades are the main way in which we redirect our eye to select particular items. Saccades are easily identifiable from a record of eye position samples, because they have a distinct velocity profile such that the eye rapidly accelerates to a peak velocity of about 500˚/s. During saccades, vision is suppressed (see Greenlee & Kimmig, this volume), and so the processing of visual information takes place largely during fixations, where the eye is relatively still. It is therefore assumed that the location and duration of fixations reflects what is being processed at a given moment in time. This assumption relies on a tight link between overt and covert attention. It also neglects the fact that saccades take at least 100ms to prepare, which means that at least part of the time during a fixation is devoted to saccade programming. However, in complex stimuli it generally makes sense to talk about attention and fixation as synonymous (consistent with the Active Vision approach: Findlay & Gilchrist, 2003).

Researchers can use the measurement of saccades and fixations in complex scenes to answer many different questions about visual attention and information processing. The top panel of Figure 1 shows the series of fixations and saccades made by a single person viewing a complex scene (despite this sort of diagram, when real

saccades are measured with high precision they are not straight and often show curvature). It is clear when we repeat this for many people looking at the same image (see Figure 1, bottom panel) that there are some consistent patterns in *where* people look. One set of questions, therefore, concerns how to represent and predict these patterns. The next section considers a widely used approach to these questions: the saliency map approach.

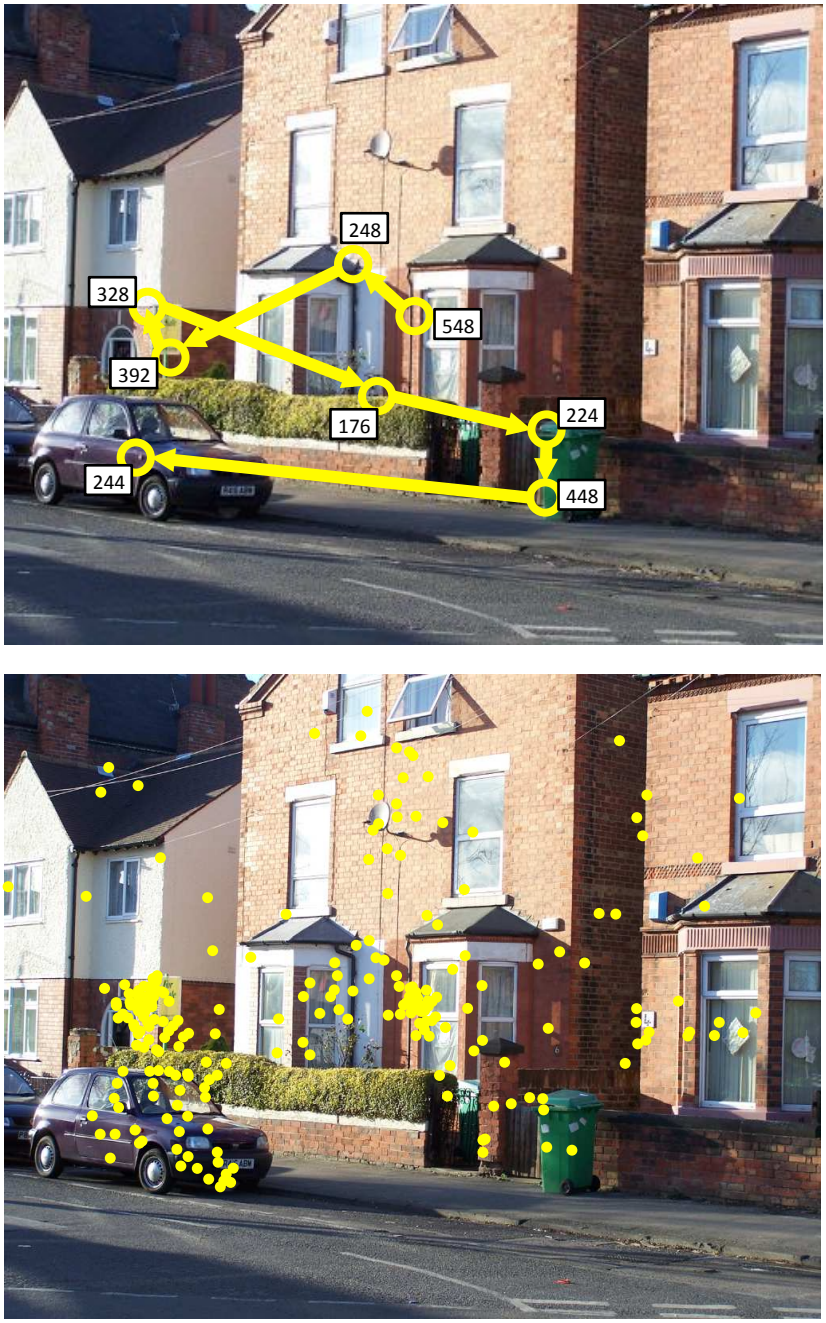


Figure 1. When someone views a scene, they make a series of eye movements (top panel). These consist of fixations (circles) and saccades (arrows). The fixations have a position, a sequential order and a duration (values in milliseconds). Combining the fixation locations across many observers reveals variability, but also clustering on certain regions (bottom panel).

A related but distinct set of questions concerns the *order* in which people attend to different elements of a picture. Investigating this order requires methods for comparing and manipulating sequential "scanpaths", the topic of both classic and contemporary research which is discussed in the second half of the chapter.

4. Historical Annotations

Although saccades and the "path" followed through an image were familiar to early researchers from introspection and observation, it was not until the 20th century that eyetrackers were used to measure this with any precision. Researchers often chose to focus on simpler and more controlled experiments, and it was only later that improvements in technology and computer vision techniques led to more experiments with complex images.

Dodge and Cline (1901) are often credited with developing the first optical eyetracker, which laid the foundation for tracking based on corneal reflections.

Buswell (1935) used an improved eyetracker at the University of Chicago to measure fixations and saccades in two-dimensional images. His observations helped to define the enduring questions of how eye movements were related to image content and viewer cognition.

Yarbus (1967) used a suction cup to record eye movements over a variety of pictures, emphasizing that the active sequence of eye movements changed with the viewer's task.

Noton and Stark (1971) defined the term "scanpath" and incorporated these fixation sequences into a detailed computational model which foreshadowed the later focus on

eye movements and embodiment (which links perception to physical and motor states).

Mackworth and Thomas (1962) developed one of the first mobile eyetrackers. Mackworth went on to help conduct several early and important studies on eye movements in scene perception (Loftus & Mackworth, 1978; Mackworth & Morandi, 1968).

Land (1993) conducted several pioneering experiments using mobile eyetracking which placed a new emphasis on the importance of action beyond just "looking at pictures".

Itti, Koch and Neibur (1998) and Itti and Koch (2000) presented a fully implemented model of visual saliency which could be applied to arbitrary images and tested with human fixation data.

Thanks to advances in eyetracking technology, computing and image processing, it has never been easier to measure fixation locations from an observer viewing a digital image. Modern research investigating how people look at pictures uses a range of devices, measures and models. This chapter reviews two general frameworks for analyzing and explaining the resulting data (saliency maps and scanpaths), and I begin the next section with a review of how we can quantify looking behaviour.

5. Spatial analysis of fixations: Saliency and saliency maps

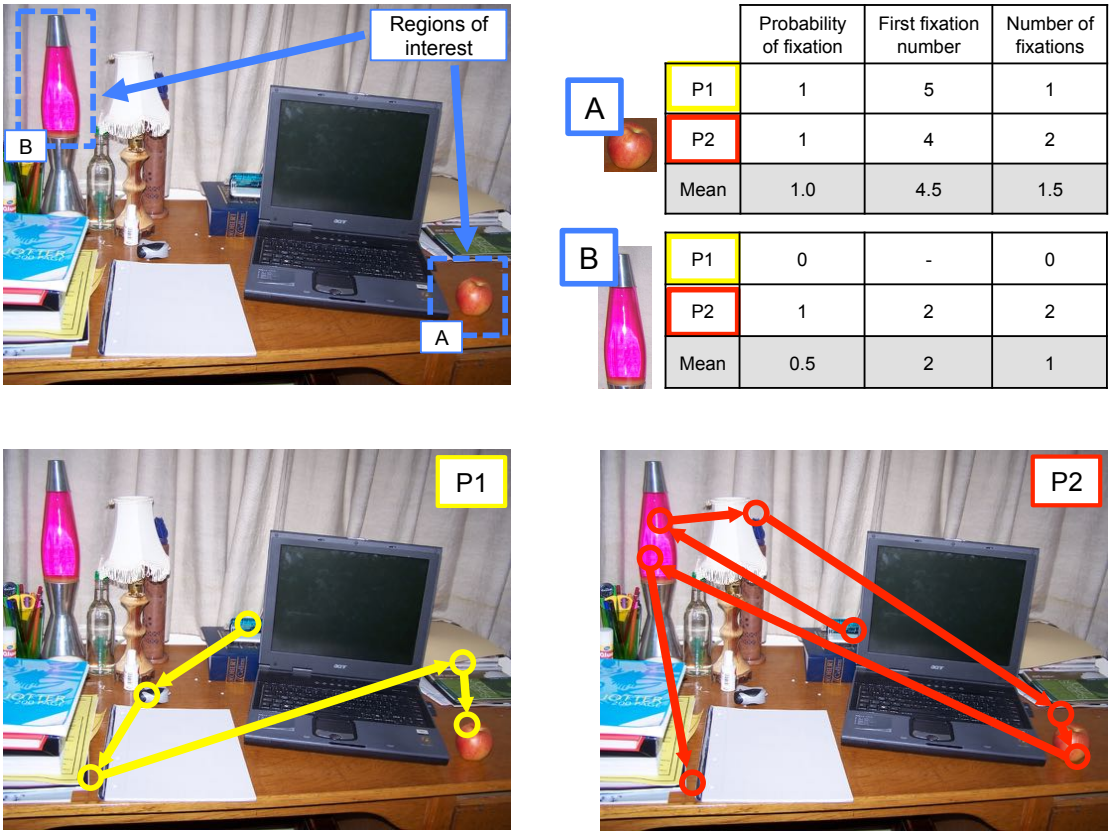
Answering the question "*Where* do people fixate?"

This section considers a particularly influential way of representing and explaining where people look in complex stimuli, with reference to the features in the image. Before discussing this approach in detail, it is worth specifying some of the different measures that researchers typically use to quantify where participants look. Table 1 describes some of these measures, their definitions and the way that they are commonly interpreted. These measures are not exhaustive, and different terms are sometimes used for the same measure. This means that it is important for researchers to clearly define their dependent variables when reporting results.

Measure	Definition	Interpretation
Trial level measures		
Average saccade amplitude	The mean (or sometimes median) amplitude of all saccades made, in degrees of visual angle	Greater values = larger shifts between points of interest in a given scene
Fixation dispersion or "spread"	The standard deviation of all *x* and/or *y* fixation coordinates, in degrees of visual angle	Greater values = fixations which are more spread out in space
Region of interest measures		
Probability of fixation	A binary, yes/no variable representing whether or not a region has been fixated at least once in a viewing episode	Regions which are fixated have received more attention than those which are not
First fixation time / number	The time / ordinal fixation number at which the first fixation on a region occurs	Lower values = region which is fixated earlier and prioritised by attention
Number of fixations (on a region)	The total number of fixations landing within a region of interest	Greater values = more interest or attention devoted to this region
First fixation duration	The duration of the first fixation on a region of interest	Greater values = more extensive or elaborate processing of the region on the first look
First gaze duration	The sum duration of all fixations made on a region of interest on the first pass (i.e., before exiting this region)	Greater values = more extensive or elaborate processing of the region on the first look
Total gaze duration / inspection time	The sum duration of all fixations made on a region of interest, including refixations	Greater values = more extensive or elaborate processing of the region over an extended period

Table 1. Some of the most commonly used measures for quantifying where people look in images.

The measures in Table 1 can be aggregated to give a summary description of the fixation patterns from many participants and trials. Most modern eyetrackers come with software for automatically calculating these statistics. The process for deriving these measures is straightforward and consists of allocating each fixation to a region of interest, based on the (x, y) screen coordinates of the fixation. The region of interest might be defined by a bounding box (most easily, a rectangle), or by a circle with a given centre and radius (thus including all fixations within a certain distance of that point). For example, Figure 2 shows one of the stimuli used in Foulsham and Underwood (2007). In this picture, we identified two regions of interest: a key object (a piece of fruit) and a region containing the brightest, most salient features in the image. Figure 2 shows how three measures can be calculated for each person viewing this image (see also Table 1). The probability of fixation merely records whether a given area of interest has been fixated or not (1 or 0). When averaged across the two example participants, region A (the apple) has a fixation probability of 1.0 because it was fixated by 100% of the observers. Region B (the lamp) has a fixation probability of 0.5 because only half of the observers looked at this region. The other measures in Figure 2 show that region A is first inspected after 4.5 fixations, and attracted 1.5 fixations, on average. In Foulsham and Underwood (2007), the same regions were present across many pictures in the experiment, and because all participants viewed all these pictures, measures could be averaged across images and then across participants to derive a description of the general tendency for people to look at either of these key areas.



		Probability of fixation	First fixation number	Number of fixations
A	P1	1	5	1
	P2	1	4	2
	Mean	1.0	4.5	1.5
B	P1	0	-	0
	P2	1	2	2
	Mean	0.5	2	1

Figure 2. A region of interest analysis with two regions (*A* and *B*, see top left panel) and example eye movements from two participants, *P1* and *P2* (bottom panels). The top right panel shows some statistics for each region and participant. For example, region B is fixated by P2 but not by P1, giving a (mean) fixation probability of 0.5.

One might well ask why so many measures are necessary to represent where people look. In some cases, several measures will yield the same conclusions. This may be because the measures are formally identical (e.g., the total gaze duration and the sum of all the fixation durations); because they are logically related (e.g., making a greater number of fixations must also mean making a greater number of saccades); or because they are merely correlated (e.g., it is normally, but not always, the case that a greater number of fixations corresponds to a longer total gaze duration).

It is important to remember that whichever measures are used, they are attempting to condense an (often complex) set of spatiotemporal patterns. The actual measures to analyse, which will reflect the hypotheses of the particular study, should be chosen with care. Figure 3 depicts some of the questions that researchers should ask themselves when choosing which measures to use. Referring back to Figure 2, we can see some of the subtleties in using these statistics to compare regions of interest. In one case, there is a region which is fixated by both participants (region A). How early and how frequently this region is inspected can be quantified by calculating the first fixation number and the number of fixations on this item. However, these measures may not make sense if a region is rarely fixated. Calculating the first fixation number is only possible if the region has been inspected at least once (and so there is a missing value for participant P1 looking at region B, see Figure 2). Here it might be more useful to quantify the region fixation probability.

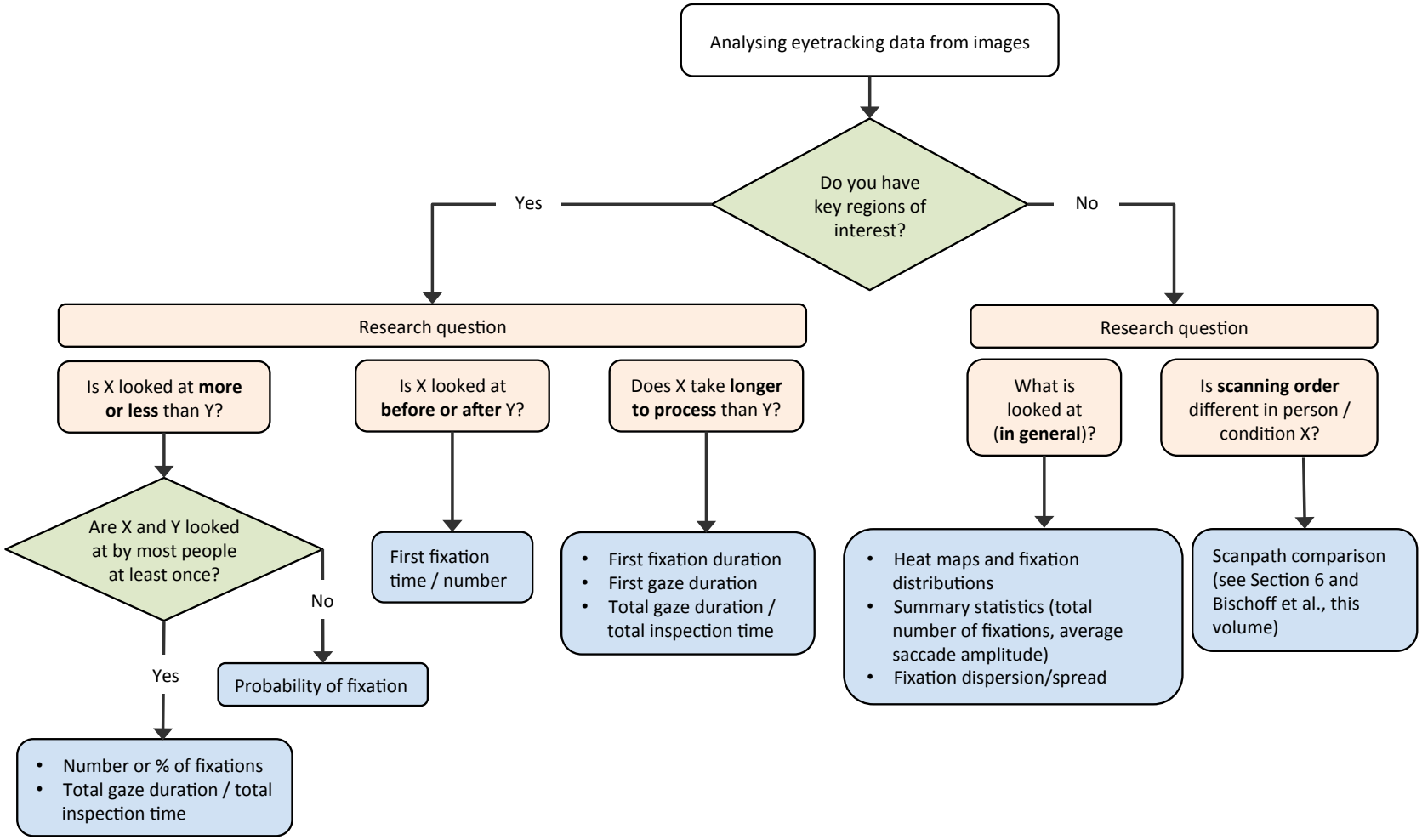


Figure 3. A flowchart depicting some of the ways in which researchers can select eye movement measures that answer their research question.

Now that we have discussed some measures of how much a particular region or object is attended to, we can consider how to explain this in terms of a representation of the features in the scene: a saliency map. This approach has been highly influential because it allows researchers to analyse complex natural scenes and relate them to eye movements in a principled way. The underlying image processing techniques are often freely available and relatively easy to use, which means that investigators in a range of fields can produce predictions for their stimuli. For example, users in marketing might wish to determine the saliency of an advertisement before measuring how often observers look at this item. Alternatively, a researcher might want to measure saliency so that they can control for the visual properties of two regions of interest in an experiment (meaning that any differences in how these are inspected must be due to cognitive factors). In the next sections, I discuss the theoretical and methodological background for the saliency map approach. This involves background work in the psychophysics and computational neuroscience of attention. Much of this material is technical, and there remain debates about how best to compute saliency, and whether it actually tells us anything about fixations. However, many of the concepts and methods involved will apply to any spatial model of eye movements in scenes.

The concept of a saliency map is founded on classic theories of attention

The term "saliency map" (or "salience map") has its roots in attempts to produce a computational model of visuospatial attention. Based on experiments

investigating visual search, and under the framework of feature integration theory, Treisman and Gelade (1980) described the process by which an observer can select a single object amongst an array. When simple features define the target, such as when one has to find a red square among many blue squares, it "pops out" and is found very easily. It is straightforward to imagine a control mechanism which could code for the colour at each location, and filter out only the region where colour = "RED". But how might such a mechanism be implemented in the case of more complex "conjunction" targets, such as a red square amongst red circles and blue squares? The solution in Treisman's model was a "master map" which combined the different basic features (colours, orientations and intensities) which were present at each location. The effortful shifting of attention while looking for the target is then determined by scanning of the master map.

A master map which combines multiple features into an abstract representation of attentional priority has been a fixture in subsequent work on attention and search. In Wolfe's (1994) Guided Search model, an "activation map" carries out the same function, prioritising (i.e., ranking) those locations in a search task that are most likely to contain the target. The activation map is therefore the mechanism by which preattentive, parallel processing of basic features is combined with top-down control in order to guide attention. Koch and Ulman (1985), meanwhile, called the combined representation guiding attention a "saliency map". Koch and Ulman's conceptual paper described the saliency map as a topographical representation which could explain selective attention in a way that was plausible given primate neurophysiology. Importantly, the saliency map was seen as an "early" visual representation which was based mostly on simple visual features. The focus of

attention was then determined by a "winner-take-all" process which selects the most salient location.

The brain may represent saliency

The topographical organisation of neurons in the early visual system was well known to those theorizing about attention in the 1980s. Single cell recording in cortical areas such as V1 showed that neurons were highly spatially selective, coding for particular features, and providing the foundation for a set of basic feature maps. However, it subsequently became clear that elsewhere in the brain there are cells which respond differently to the same stimuli, depending on the current attentional priority. For example, the superior colliculus is crucially involved in the control of eye movements. The activity of cells in this part of the midbrain does not just depend on whether there is a visual stimulus in their receptive field. Instead, cell responses are enhanced when observers are planning to make a saccade to this stimulus (Goldberg and Wurtz, 1972). Thus, deploying attention and eye movements to a certain location increases activity in colliculur neurons coding for this location.

Beyond the general attentional modulation of visual responses, there is mounting evidence that frontal and parietal areas involved in the control of eye movements can be thought of as implementing a saliency map (see Treue, 2003). For example, microstimulation of the frontal eye fields is associated with increased responding in spatially selective regions of V4 (Moore & Armstrong, 2003). Thus, the neural activity integrates relevance—signified by the process of preparing an eye movement—with visual distinctiveness. Modern neuroscience has provided

increasing evidence for spatial priority maps in frontal and parietal areas of the brain, as well as increasingly sophisticated discussions about how these are involved in the guidance of behaviour more generally (see Bisley & Goldberg, 2010; Zelinsky & Bisley, 2015, for recent reviews).

It should be clear from this background research that the term saliency map has been used in multiple, overlapping ways: as an abstract, master map for attentional priority; as a neural mechanism for combining visual activity; as a bottom-up predictor of where people will look; and as any heat map type representation of fixations (see Textbox 1). These terms have not always been applied consistently, and so it is important for researchers to provide a precise definition. However, the main focus of this section will be a particular style of computational model of visual saliency which has been widely used, and which provides an estimate of the bottom-up feature contrast in complex stimuli.

Textbox 1: Heat maps and fixation distributions

It is often useful to represent the distribution of fixations across an image by presenting a heat map: a spatial density plot showing how frequently each part of the picture has been inspected. The discrete fixations are first transformed into a continuous distribution, often by convolving a binary map with a symmetrical Gaussian function. This can also be thought of as iteratively adding a "blob" at the location of each fixation. The standard deviation (σ) of the Gaussian affects the granularity of the heat map (Fig. A), and should be chosen to reflect error in eye position and the size of the fovea (e.g., 1 degree of visual angle). The result can be plotted as an "attentional landscape" with "peaks" or "hotspots" showing the places which are fixated most often.

There are a number of somewhat arbitrary factors that can be changed when making such maps. As well as σ, some researchers represent fixation duration (by scaling the height of the Gaussian), and some may produce heat maps for each participant which are then averaged. Wooding (1995) and Le Meur and Baccino (2013) describe heatmaps in more detail.

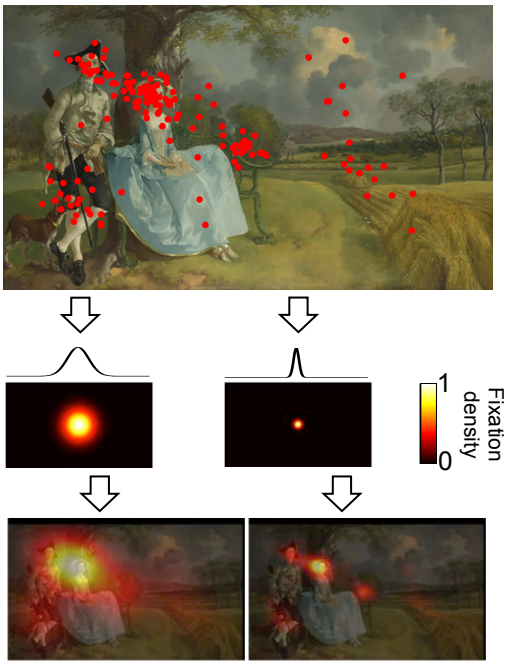


Fig. A. The same fixations, convolved with two differently sized functions. The resulting heatmaps are plotted over the original image.

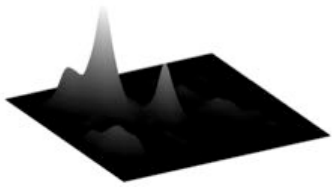


Fig. B. An attentional landscape from the same fixations in Fig A.

The Itti and Koch saliency map model

Why is it that some regions are inspected more than others? One possibility is that there might be a set of visual features, which can be identified before the planning of an eye movement, and which signal those regions that should be prioritised for attention and fixation. A likely candidate feature is *contrast*, in the general sense that something which stands out from its background might be worth looking at, just as a feature search target pops out in simple visual search. This idea underlies the concept of visual saliency that was explicitly modelled by Itti, Koch and Neibur (1998), and then applied to eye movements by Itti and Koch (2000, 2001).

Itti and Koch built on the work of Koch and Ulman (1985), who were inspired by both human psychophysics and primate physiology, and implemented a computational model which combined bottom-up visual features using image processing. Unlike the simple case of searching for a red square amongst blue squares, the problem of extracting and combining features from a natural scene quickly becomes complex. Itti and Koch (2000) proposed a series of steps, implemented and released as a programming toolkit, which could take any arbitrary digital image as input (see Figure 4 for an example, and http://ilab.usc.edu for downloads and software).

The first step extracts basic visual features at a range of spatial scales. Each feature is associated with a map, with the value at each point in the map representing the presence of that feature at that location. In static images, colour, intensity and orientation features are extracted using spatial filters (whereas flicker and motion channels are also available in dynamic stimuli). For example, the intensity map simply represents the amount of light at each point in the image, with black regions

having low intensity and very bright regions having high intensity. Colour features are combined to compute colour opponency, while the orientation channel is assembled from edge detectors of different orientations. In order to identify such features at both coarse and fine scales, maps are derived from progressively sub-sampled versions of the image.

The second step is to combine the resulting scaled feature maps in a way which highlights feature contrast. This is accomplished through a centre-surround arrangement which pits fine and coarse feature maps against each other. The result is a "conspicuity map" for each feature, where the locations with the strongest activation are those with features which stand out from the surrounding background. An important consideration for the particular computational implementation at this point is how to combine different features. Should something colourful be given the same priority as something which is brighter than it's background? Although several solutions to this are possible, in Itti & Koch's implementation there is normalisation and between-feature competition, such that if there is greater contrast in one feature dimension then it will be emphasised at the expense of features with lower contrast. After normalisation, the conspicuity maps are added together to give an overall saliency map, with "hotspots" showing the most salient regions with the highest feature contrast.

The last step of the model, which makes it particularly suited to applying to eye movement research, is that it uses the saliency map to make explicit predictions about the locations which will be selected by covert and overt attention. Attention moves to the most salient location via a winner-take-all network. The saliency of this location is then suppressed (in a way similar to "inhibition of return", the mechanism proposed by Posner et al., 1985, for explaining delayed orienting to previously

attended locations). This allows the focus of attention to shift to the next most salient region.

The Itti and Koch saliency model has been very widely cited and applied to many different problems in human vision and computer science (see Borji & Itti, 2013, for a recent review). Its success can be attributed to the fact that it produces real, tractable predictions for any arbitrary visual stimulus, with the algorithms for producing these predictions freely available. I will now describe some of the practical considerations involved in computing a saliency map, before discussing the experiments that have been carried out to test this model with human eye movements.

At this point it is important to note that there are now many different "saliency map models". In the almost 20 years since the original Itti and Koch (2000) model was proposed, it has been regularly revised and improved. Several different implementations have been released, which may differ in both algorithmic detail and actual predictions. Other researchers have proposed different underlying features, or tried to incorporate aspects such as depth or motion. In many cases it is also possible to incorporate some "top-down" modulation, where the model learns or is given information about which features are important (for example by adding a face detector, or training a model to look for a certain object). All of this means that it is important for researchers to be specific about what model and implementation is being used, and with which settings.

Table 2 describes some of the related, bottom-up models which have been published, with particular emphasis on those which are available to download (for a much more exhaustive list, see Borji & Itti, 2013; For model evaluations, see Kumerer et al., 2015). Critically, many of the methods I will describe next can be applied to any of these models. Moreover, many of the principled criticisms that have

been levied at the saliency map approach are problematic for *any* model which is based solely on bottom-up features.

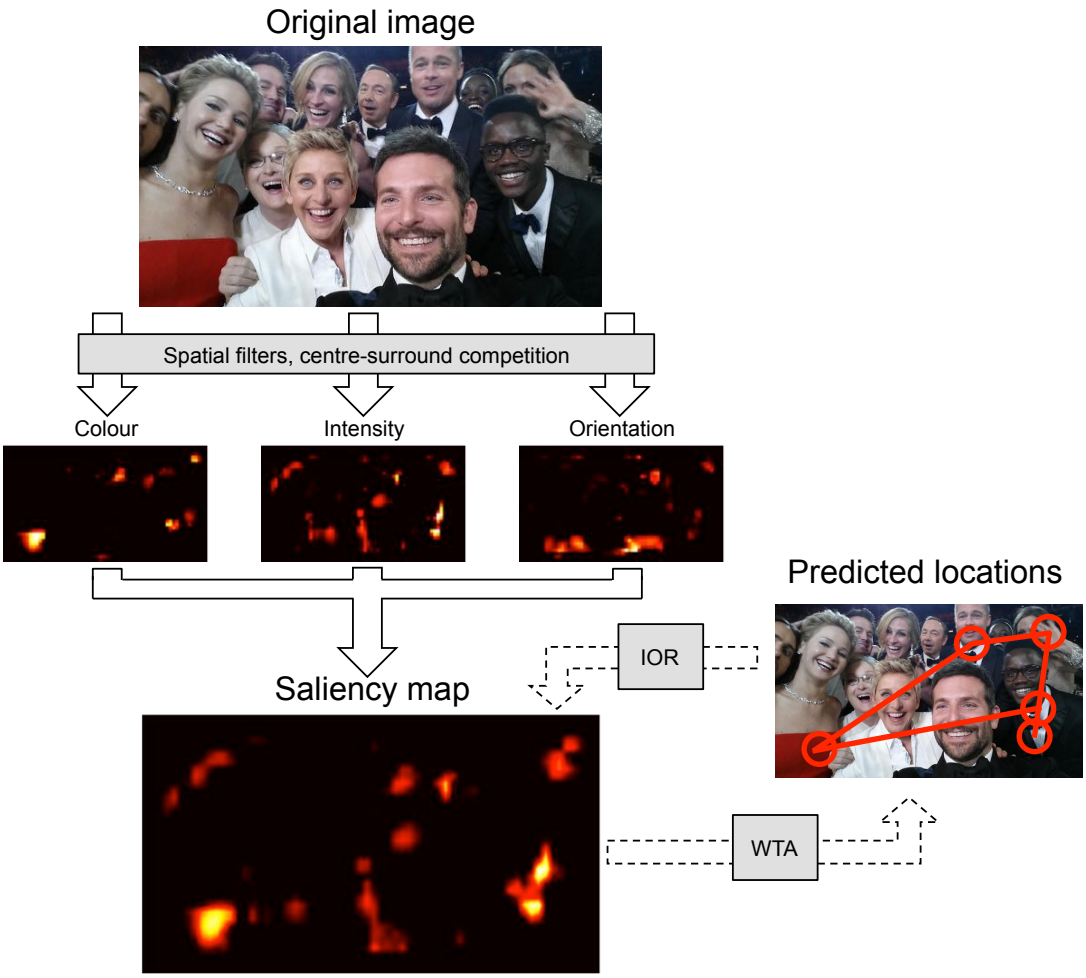


Figure 4. An example of applying the Itti and Koch (2000) saliency map model to a complex image, the "most-tweeted photo ever" from the 2014 Academy Awards. The image is analysed within three different feature channels which are combined in a centre-surround fashion to highlight feature contrast. The resulting saliency map predicts a series of attended locations through a winner-take-all (WTA) process with inhibition of return (IOR). Salient locations include the black-and-white contrast of a tuxedo and a conspicuous red dress.

Model	Noteworthy features	URL
Itti et al., (1998; Itti & Koch, 2000) saliency map model	Original model implemented in C++ as part of the iNVT toolkit	http://ilab.usc.edu/toolkit/
Walther & Koch (2006) saliency toolbox	Implements Itti model but also aims to identify and parse "proto objects"	http://www.saliencytoolbox.net
Harel et al., (2006) graph-based visual saliency	Uses graphical models to identify conspicuous regions. Code also includes Itti model.	http://www.vision.caltech.edu/~harel/share/gbvs.php
Itti & Baldi (2006) Bayesian model of surprise	Defines saliency mathematically, according to change in prior beliefs. Tested with eye movements in video.	http://ilab.usc.edu/surprise/
Bruce & Tsotsos (2005) AIM: Attention based on Information Maximization	Uses measure from information theory to define salient regions. Applied to visual search (with top-down information)	http://www.cs.umanitoba.ca/~bruce/datacode.html
Le Meur et al., (2006)	Using Itti model as a baseline, introduces more detailed stages to mimic human visual system.	Code not readily available, but see http://people.irisa.fr/Olivier.Le_Meur/
Zhang et al., (2009) SUN: Saliency Under Natural statistics	Defines local saliency in Bayesian terms. Combines this with top-down information.	http://cseweb.ucsd.edu/~l6zhang/code/imagesaliency.zip
Judd et al., (2009)	Introduces a widely-used eyetracking dataset and uses machine learning to identify low-level features at fixation (as well as mid- and high- level features such as horizon and face detectors).	http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
Vig et al., (2014) Ensembles of Deep Networks	Uses "deep learning" to learn the optimal bottom-up features for fixation. Currently one of the best performing models (Kummerer et al., 2015).	http://coxlab.org/saliency/

Table 2. A selection of "saliency models" which build on the ideas from Itti and Koch (2000).

Guidelines for eye movement researchers computing a saliency map

It is clear from Table 2 that there are a bewildering number of models available, and this is not the place for considering them all in detail. One of the major contributions of the modelling community working on saliency maps is that there is a large amount of open data and code available. I recommend that interested readers try out some of the software available. Although the original version requires some knowledge of command line programming and C++, several user-friendly toolboxes have also been developed. For example, Figure 4 was created using Version 3.0 of the Saliency Toolbox (Walther & Koch, 2005; 2006). The Graph-Based Visual Saliency implementation by Harel et al., (2006) also provides an implementation of saliency which is relatively easy for beginners to try. Both toolboxes run within MATLAB (Mathworks) and provide a set of functions, documentation and a graphical user interface. In this section I will offer some brief guidelines for eye movement researchers wishing to use such functions with complex images.

The first step in generating saliency map predictions is to load in a digital image for the stimulus. One way of thinking about this simulation is that the model should receive the same visual input as the human observer. Thus the image used should be the same as that seen by participants. The size of the image is one of many things which can change the specific output of the model, as well as affecting the length of time taken for the simulation to be complete. Some models may treat grayscale and colour images differently, and it is useful to know how the images are encoded (e.g., as an RGB image).

Next, the image is passed to the saliency functions for analysis. In some cases the processing involved is considerable, meaning that it can take some time.

Typically, model output includes a continuous map of the image, where each pixel represents the saliency of the corresponding point in the image. This map can be displayed as a heatmap. However, it is important to be clear how this map is scaled, and the colours that are used in displaying it (for example, how is the minimum or maximum value in the map displayed?).

It is particularly important to understand how the images are scaled so that one can compare the saliency map output to objects in the image or fixated locations (see next section). It is also important to understand that the model essentially pits different parts of an image against each other, calculating the *relative* saliency of each region. Thus, if the research question requires comparing between different objects, both objects need to be present in the same image. It may not be straightforward to compare the saliency maps of two different images, particularly if they show very different variances in terms of the features present and their spatial distribution.

This description was written as a practical guide for people who want to produce saliency-based predictions for a particular image. However, as we shall see in subsequent sections, the saliency map model is not without its critics. Moreover, it is regrettable that there are multiple free parameters which can be changed, many of which are not specified by authors using this software, which may lead to problems replicating the results. It is therefore useful for researchers to experiment with different settings and be clear which differences in saliency are robust and which are highly sensitive to changes in model parameters.

Testing the saliency map model: Multiple methods, none perfect

The Itti and Koch (2000) model was billed as a mechanism for shifts of overt and covert attention, making it ideal for producing bottom-up predictions for eye movements in complex stimuli. In this section, I will describe some of the ways in which these predictions have been tested. To begin, we should bear in mind that the model produces both a continuous spatiotopic map, representing the saliency of each location, and a simulated series of attention shifts. Both of these outputs can be used to test the model directly, and these tests fall into three main groups.

The first type of analysis examines the strength of the activation in a feature or saliency map at each fixated location. If the saliency map is a good prediction of the regions which will be inspected, then fixations will select the locations with high values, and avoid those locations where saliency is low. To perform this analysis, we can take the saliency value from each fixated location and average it across multiple fixations and participants. The saliency map is often represented at a lower spatial scale (i.e., it is smaller and represents regions more coarsely than the original image), and so locations will need to be scaled appropriately so that the correspondence between points in the image and on the map is maintained. This is normally just a case of determining the size of the saliency map relative to the image and then scaling position coordinates accordingly. Rather than relying on exact fixation coordinates, one could also take all the values from a small region around fixation, perhaps within 1 degree of visual angle, compensating for potential errors in eye position measurement, and for the fact that the fovea takes in information from an extended area. These values could be averaged (or summed), or the maximum value could be taken. When evaluating saliency values, it is important to consider the way the map is scaled or normalized (e.g., what is the minimum and maximum value in the map, and how distributed are these values), and differences in the distribution of saliency can

make it difficult to compare between different images. If the saliency map is conceptualized as a probability distribution, then all the values will be positive and sum to 1. Other implementations will have a fixed range (e.g., between 0 and 1, or between 1 and 255 which is common in 8-bit digital images). One way to take these into account is to calculate what Peters et al., (2005) dub the normalized scanpath salience (NSS; see Figure 5A). To compute this measure, the saliency map is normalized by subtracting the mean saliency across all locations and dividing by the standard deviation of saliency values. This produces a z-score, and thus shows how many standard deviations a particular location is above chance. Using a standard saliency model to predict fixations in outdoor images, Peters et al. reported a mean NSS of 0.69 which was far greater than that expected by chance (an NSS of 0).

The second type of analysis compares the overall distribution of many fixations with the saliency map, in a manner similar to a correlation. If the model is predicting where people look, then there should be a positive relationship between saliency and fixation density. Le Meur and Baccino (2013) give a useful summary of some of the steps and metrics used in this type of analysis, and they also provide some computer code for these analyses. One initial approach is to convert a list of fixation locations (e.g., those from multiple participants viewing over an extended period of time) into a continuous fixation density distribution. Such distributions can be represented as heatmaps, showing the relative frequency with which different regions are inspected (refer back to Textbox 1). Comparing a saliency map and a fixation distribution can be as simple as calculating a Pearson correlation coefficient between the two. If points with high saliency are also locations with a high density of fixations then the correlation between the two distributions will be positive. However, because the distributions involved may violate parametric assumptions, it is preferable to use a

non-parametric method for comparing the probability distributions. One such metric is the Kullback-Liebler (K-L) divergence. The K-L divergence is a measure from information theory which quantifies the difference between two probability distributions. The result is a score—the number of bits—indicating how different the two distributions are, with a score of 0 indicating identical distributions. A better match between a saliency map and a fixation distribution would give a lower K-L divergence. This metric is discussed in detail by Tatler et al., (2005), who were also among the first to use the signal detection methods which have become standard in those evaluating saliency models.

The signal detection approach uses the saliency map to discriminate between fixated and non-fixated locations, by applying a threshold. Locations with saliency higher than the threshold are classified as fixated points, and the threshold is gradually increased, allowing the hits (correctly identified fixation locations) and false alarms (non-fixated locations classified as fixated) to be tallied. For example, at a low threshold many locations will be selected, leading to many false alarms (as well as some hits). At a high threshold, only the most salient locations will be classified as fixated, and if the saliency map is a very good predictor of fixation these will all be hits. Using the rates of hits and false alarms at each threshold, a receiver operating characteristics (ROC) curve is plotted, and the area under this curve (AUC) quantifies how well the saliency map can discriminate fixated locations (see Figure 5B). This method is robust to differences in the distribution of saliency in fixated and non-fixated locations, and does not rely on parametric assumptions. Critically, because the ROC method is based on the ranks of all the points in the map (i.e., the point with the highest saliency, followed by the next highest, and so on), it is not affected by any monotonic scaling of the actual values. A saliency map which provides no

information about fixated locations will lead to an AUC of 0.5. Another advantage of this method, widely used in computer vision, is that it can de-confound effect size and statistical significance. For example, Tatler et al., (2005) report an AUC of 0.57 for the predictiveness of a luminance saliency map. This value was statistically very different from chance (according to bootstrapped confidence intervals), but it also leaves much of the difference between fixated and non-fixated locations unexplained (57% is by no means an impressively large effect).

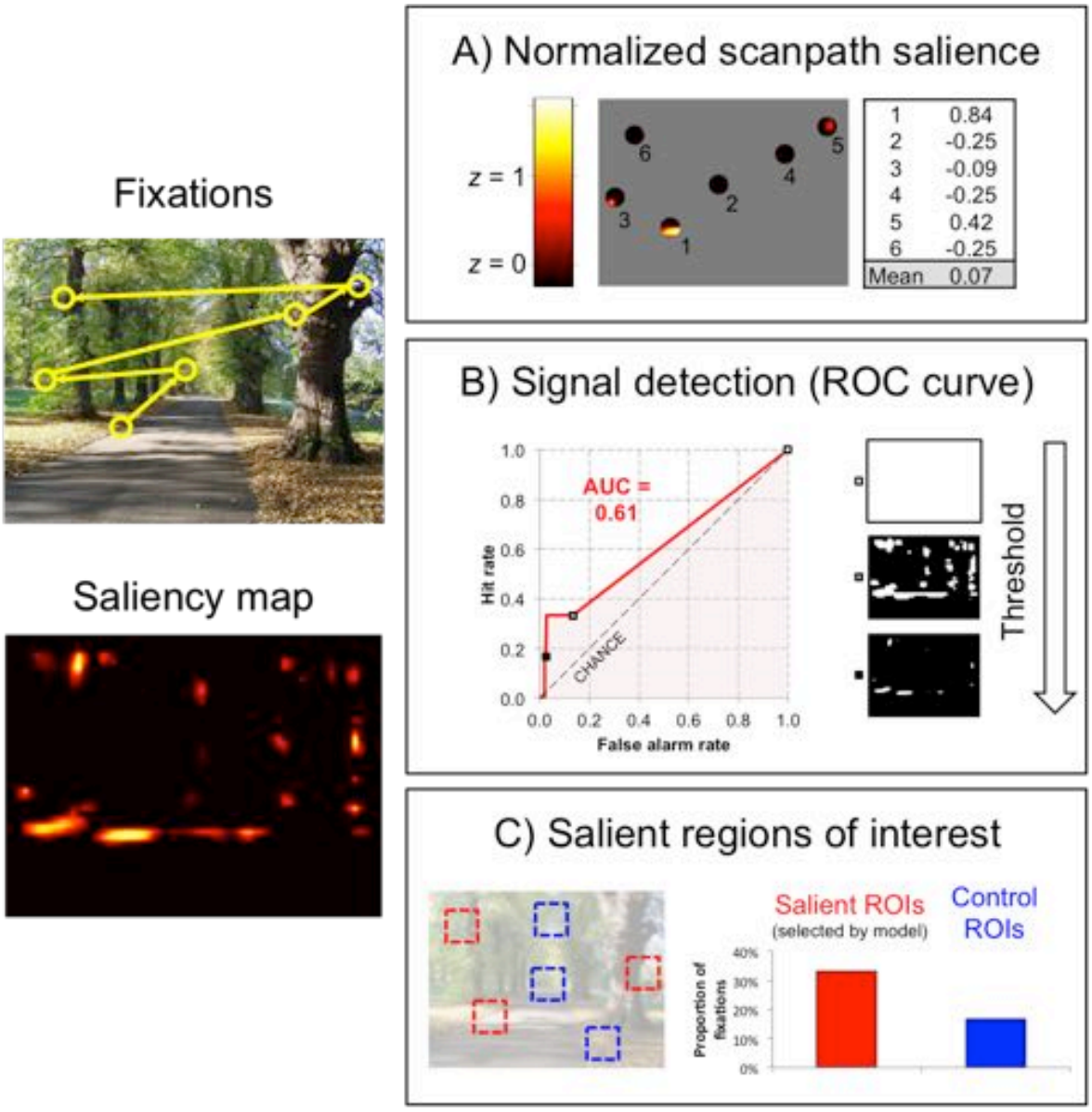


Figure 5. An example of three different ways to compare fixations on a scene to a saliency map. In A), we take the average of the z-transformed map value at each location. In B), an ROC curve is plotted by applying a variable threshold to the map and observing the correctly predicted fixated locations (hit rate) and the false positives. Alternatively, salient regions of interest can be identified from those areas selected by the model, and compared to areas that are not (C). The methods are applied to the six example fixations in each case.

These general analyses are useful for comparing fixations and saliency over an entire image in a theory-neutral way. However, often researchers are interested in certain objects or regions which have a particular significance based on theory or application. The third type of analysis uses the sequence of saliency-predicted shifts of attention to identify key regions in the image, and then determines how often these regions are fixated. For example, in Figure 2 I described an experiment where "target" objects were made more or less salient by placing them in different visual settings. The relative saliency of target objects was determined via the model, according to the number of simulated shifts of attention that were made before they were selected. For example, one might classify a "high saliency" object as one which is selected by the model very early (on the first or second simulated fixation). Such objects could be compared to "low saliency" or "control" objects which were not selected by the model after ten shifts of attention. This approach has a number of advantages. First, it tests one of the key strengths of the saliency map model, as proposed by Itti and Koch (2000), which is that it predicts an actual sequence of fixations and not just a continuous map. Second, it makes it possible to compare between targets which have been matched in other ways (such as according to their semantic meaning by virtue of being the same type of object). As we shall see, the failure of correlational approaches to take semantics into account has led to difficulties in evaluating the saliency map model. In Foulsham and Underwood (2007), we found that in the absence of a strongly constrained task, objects (pieces of fruit) which were more salient according to the model were more likely to be fixated and were fixated earlier. Figure 5C gives another example of comparing the fixations on salient and non-salient regions.

When evaluating these analyses, we often have to compare the results to some kind of null hypothesis, indicating the relationship that we would expect if saliency

does not predict fixation. For example, the NSS compares saliency at fixation with the average saliency across the whole image, and the AUC compares fixated locations with all other non-fixated locations in the image. However, because neither fixations nor saliency are uniformly distributed, comparing against a "chance" distribution which samples uniformly across locations is problematic. The issue is that such a comparison assumes that all parts of an image are equally likely to be fixated. As we shall see, this may have caused the role of saliency to be overestimated because fixations in the centre of an image are credited to salient features when in fact this is a generic spatial bias which is manifested regardless of the scene. Eye movements across many different images are systematically biased to particular spatial locations in a variety of ways (see Textbox 2). Whichever metric is used for comparing saliency models and fixations, the best approach is to select a comparison distribution which reflects the general spatial biases inherent in eye movements across images. For example, rather than comparing saliency at fixated locations to the average across the whole scene, they can be compared to values from a "shuffled" dataset which uses positions selected by human fixations in other images. This shuffled dataset thus reproduces the image-general spatial biases, ensuring that only predicting fixation patterns on a specific image is credited to the saliency or feature-based model. Alternatively, one can use a generic, non-uniform comparison distribution (such as that recommended by Clarke & Tatler, 2014, which models the general central bias seen in image viewing).

Textbox 2: Systematic patterns in scene viewing are not random

Tatler and Vincent (2009) showed that we can predict fixation locations just by knowing how the eyes move in general. This is because, regardless of the image, observers show systematic biases towards certain locations and saccades. It is therefore important to investigate how these biases are related to image content and, if they are not, why they arise.

There is a strong central bias in fixations on a screen (Fig. A). This is exacerbated by the practice of cueing participants with a central fixation cross, but may also reflect the eyes' "orbital reserve". Photographers also often place objects and items of interest in the centre.

Saccades are also biased, occurring most often in horizontal directions (Fig. B). This is true even in square images, and follows perception of the layout when the scene is rotated (Foulsham et al., 2008). There is also a small but consistent bias to make an initial saccade to the left (Fig. C).

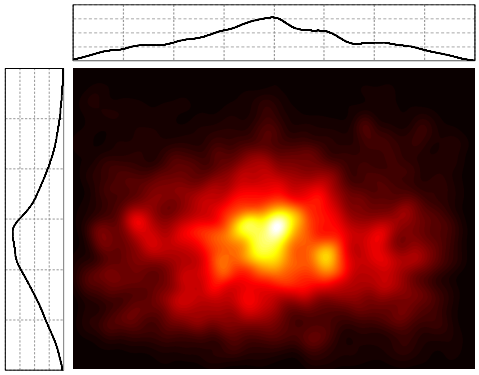


Fig. A. Heat map and histograms showing the relative frequency of fixations across scene space.

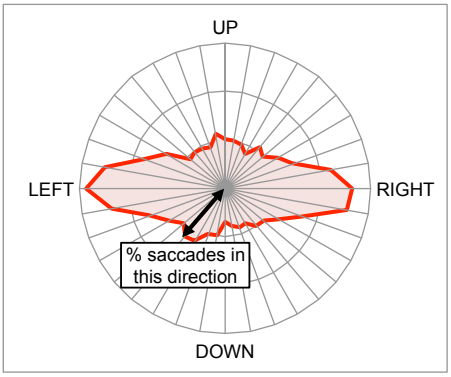


Fig. B. Relative frequency of saccades in each direction (data in Figs A and B from Foulsham & Underwood, 2008).

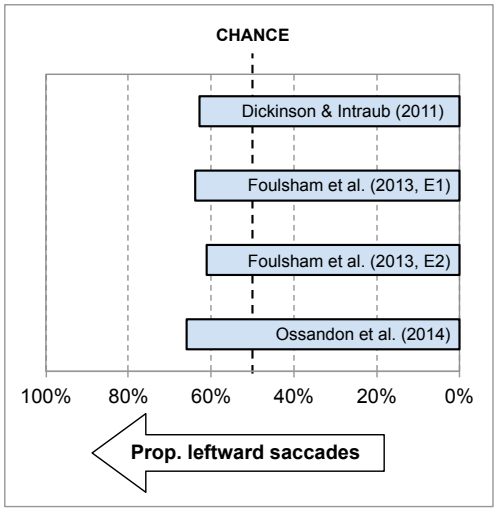


Fig. C. Recent studies reporting a greater than average proportion of leftward initial saccades.

Bottom-up features play only a minor role in the guidance of eye movements

Now that we have discussed methods for testing the predictiveness of the saliency map model, how should we judge its success? When doing this, it is useful to return to the original paper by Itti and Koch (2000), who defined their model as "bottom-up". Although the distinction between bottom-up and top-down is not always clear-cut, their model was meant to simulate early visual processes, and it contained no information about the meaning or relevance of scene regions. A glance back at Figure 4 will convince the reader that the salient items in the "Oscars selfie" image are not necessarily the most interesting, and neither are they the parts of the image which one intuits will catch viewers' attention, because the model knows nothing about faces or celebrities. Also, because the traditional saliency map approach produces the same predictions regardless of what the viewer is doing, it cannot account for any differences in behaviour based on task. These facts were noted by Itti and Koch (2000), who acknowledged that it "might be that top-down influences play a significant role in the deployment of attention in natural scenes" (p.1502), and that this was a key limitation of their model.

There does appear to be a correlation between fixation and saliency, particularly in the contrived case of a "free-viewing" task where participants are merely asked to look at an image with minimal task constraints (e.g., Parkhurst et al., 2002). We have found that saliency at fixation is higher than chance (as did Peters et al., 2005), and that highly salient objects are looked at more often than less salient objects (Foulsham & Underwood, 2007; 2008). However, it would be incorrect to conclude from this that the literature was supportive of a causal relationship between saliency and eye movements. In fact, the opposite is true, with most recent research

arguing that saliency is insufficient for explaining where we look. Some of the limitations of the saliency approach are discussed in detail by Tatler et al. (2011). The main criticisms are as follows.

First, while there may be a correlation between saliency and fixation, this correlation is really rather weak, even in ideal circumstances. In terms of the area under the ROC curve, values of around 0.6 or 0.7 are most common, meaning that the underlying image statistics provide only a modest amount of information about fixation selection. Considering that "blind" models which do not have any information about image features but only about the systematic way that people move their eyes can lead to classification performance of 0.65 (Tatler & Vincent, 2009) or 0.68 (Ehinger et al., 2009), saliency may not tell us very much extra. Whatever metric is used, it is important to benchmark it against the amount of variability that we could possibly expect to predict. Because the saliency map model is a normative model, which produces the same predictions for everyone, it obviously cannot predict individual differences between observers, and neither can it account for any random component to fixations. A common approach, therefore, is to express the predictiveness of the model as a proportion of the "inter-observer consistency" (i.e., the ease with which fixations from a given observer can be predicted by the fixations of all other observers). This reflects the fact that what we are really trying to predict is the commonalities in looking behavior across people. Peters et al., (2005) report that their baseline saliency model gives an NSS of between 39% and 57% of the inter-observer value. While this is greater than zero, it leaves considerable variance unexplained.

A second problem is that with correlation-type analyses there are numerous other factors which might co-occur with saliency and which cause regions to be

fixated. In other words, saliency may not actually cause regions to be fixated in the first place. Visually salient regions are often also semantically informative (Henderson et al., 2007) and contain meaningful objects (Elazary & Itti, 2008). Conversely, the regions which are fixated least often—consider the empty patches of sky and road in scenes such as Figure 1—are often both non-salient and not of central focal importance to the meaning of a scene. We cannot, without systematic manipulations, conclude that what makes these regions priorities to fixate is indeed their visual saliency. Although it is a matter of current debate, objects which are manually identified by humans seem to be fixated more often than predicted by saliency, and selected in a way that is more consistent with complex object features than simple edge detection (Foulsham & Kingstone, 2013a; Einhauser et al., 2008; but, see Borji et al., 2013). There are of course other meaningful objects, such as human faces, which appear to be fixated regardless of their visual saliency (e.g., Birmingham et al., 2009; See Bischoff et al., this volume).

The third major issue for saliency map models is that, when participants view images under realistic task conditions, saliency seems to play very little role and can be immediately overridden. In particular, in visual search, salient regions are completely avoided in preference for regions which look like the target, or areas where it is likely to appear (Ehinger et al., 2009; Foulsham & Underwood, 2007; Henderson et al., 2009). When the task requires fixating a target in a non-salient region, this task is completed even on the first fixation (Einhauser et al., 2008). Modifying semantically meaningful regions to change their saliency has little or no effect on their likelihood of fixation (Nystrom & Holmqvist, 2008). Beyond picture viewing, it may be best to conceptualise the "relevance" of a to-be-fixated location as

depending, not on visual saliency, but on reward in the context of active tasks (Rothkopf et al., 2007).

As described in Table 2, there has been considerable progress in the development of bottom-up saliency models. Several datasets are freely available for people to test proposed models and evaluate the results, stimulating a competition between models (e.g., the MIT saliency benchmark: http://saliency.mit.edu). Recently, this has involved applying more powerful machine-learning or deep-learning algorithms to learn the optimum features for discriminating fixated and non-fixated regions (e.g., Vig et al., 2014). Kummerer et al. (2015) develop a principled way to compare different models and metrics by quantifying the amount of information a model provides, over and above an image-independent baseline. This paper shows that bottom-up models are becoming better at predicting where people look. On the other hand, it also estimates that the very best performing model can still only account for 34% of the information gain which should be explainable based on consistency between participants. Thus the saliency map approach is a long way from being able to explain patterns in where people look.

To summarise, then, although the saliency map model may represent a plausible and tractable way to estimate bottom-up feature contrast, it is not appropriate as a catch-all model for fixations in natural images. It remains an open question whether there are real-world situations where we prioritise saliency at the expense of relevance. Instead, researchers have focused on modeling top-down factors—knowledge of a search target or a scene, and the demands of a task—and combining these with visual saliency (Navalpakkam & Itti, 2005; Ehinger et al., 2009; Zelinsky, 2008).

6. Sequential analysis of fixations: Scanpaths and scan patterns

Temporal and sequential analysis of fixations

So far, most of this chapter has dealt with answering questions about *where* people fixate. This is a sensible thing to investigate, because the eyes can only be directed at one location at a time and our eyetrackers can measure this with high precision. However, this has often led to a separation between analyses focused on spatial looking patterns and those based on the *timing* of where people look. This is surprising given that in other domains the question of when the eyes are moved is especially important (e.g., in reading; See Hyona & Kakkinen, this volume). In this part of the chapter I will describe one way of combining spatial and temporal information: by analyzing scanpaths. First, we should consider fixation duration and other temporal measures that have been investigated in scene viewing.

Turning back to Figure 1 at the beginning of the chapter, we can see that each fixation on an image has a duration as well as an order in the scanning sequence. Typically, fixations on complex images have an average duration of around 300 ms, but they can vary considerably between different observers and different images. Very short fixations (less than 80 ms or so) are rare and, because it must take longer than this for the visual system to program the next saccade, these outliers are sometimes excluded. Understanding the causes of the variability in fixation duration (what makes some fixations longer than others) is difficult in natural scenes because we are rarely in control of precisely where someone is looking or the information available at that point. However, with the support of evidence from reading and

picture viewing it is generally assumed that longer fixations reflect more difficult, more extensive, or more effortful processing of the details at that location. For example, objects which are out-of-place are associated with longer fixation durations (Underwood & Foulsham, 2006), and fixations are longer on average when trying to remember the details of a scene than when searching around for a specific object (Mills et al., 2011). Conversely, we would not expect fixation duration to be associated with superficial visual properties of an object which do not change its meaning (and hence in Foulsham & Underwood, 2007, we found that salient objects were fixated *earlier* but not for a longer duration).

There are a number of difficulties with analyzing fixation durations for particular areas of interest in complex scenes. One issue is that, although it is normally a safe assumption that fixation duration reflects processing at the current location, in some cases prolonged fixation duration may indicate covert scanning or changes in the periphery. Another is that objects are often fixated more than once, and so it is really the cumulative time spent on an object which is a better measure of processing (see Figure 3). Multiple, consecutive fixations on an object are normally described as a "gaze" (but also as a "dwell" or a "run"), and thus the first gaze duration describes the sum duration of all fixations before moving the eyes away from a particular region of interest. The pattern of fixations and refixations may be complex. For example, participants might make a short initial fixation near an object, before making a small "corrective" saccade to a better position for a longer fixation. The interaction of fixation position on an object and fixation duration has been of recent interest for several researchers (Foulsham & Kingstone, 2013a; Nuthmann & Henderson, 2010).

A good way to introduce some control into experiments in scene perception is by using gaze-contingent displays, where parts of the scene are changed in real time and in response to eye movements. Henderson and Pierce (2008) masked the scene at a critical point during a saccade and observed the results on the following fixation (the "scene onset delay" paradigm). Surprisingly, they observed that not all fixations were affected by this change. Thus, although some fixations were under direct control in response to what the observer could currently see, others were "pre-programmed". Such evidence has been used by Nuthmann et al. (2010) to propose a model of fixation durations in scenes which is driven by the underlying saccade programming.

Participants in scene viewing experiments often view a picture for several seconds at a time. As we might expect if observers are learning about the image and changing their priorities, eye movements may change over this period. For this reason, some researchers choose to look specifically at the first few fixations, arguing that this is the point when participants are drawn to certain details for the first time. Fixations in the first 1 to 2 seconds of the viewing period tend to have a shorter average duration (and these are associated with larger average saccades; Unema et al., 2005). It has been proposed that such changes reflect a switch between "ambient" or "global" exploration and "focal" or "local" scanning. However, it is not known how ubiquitous these patterns are or how they are related to scene content.

It is clear from the research reviewed in this section that it is important to think about *when* fixations are made, as well as *where*. Rather than aggregating across a number of fixations made at different times, we can also examine the presence of particular sequential patterns in where observers look. This is discussed in the following sections.

Classic research emphasized cognitive constraints on scanning patterns

The focus of recent research on visual saliency and other image-based accounts of fixation is curious, because classic research emphasised the opposite: that where we look and the order in which we look there is determined top-down. Buswell (1935) and Yarbus (1967) studied participants' eye movements while viewing paintings and images. They found that people tended to look at semantically meaningful regions: faces, objects and details important for understanding the scene. As we discussed in the previous section, it is unlikely that a purely saliency-based account of where people look can explain the concentration of fixations on such items. Moreover, Yarbus is often credited with showing that the pattern of eye movements made depended on the task that the participant was performing (although Buswell had in fact already observed this). In particular, Yarbus showed observers the same painting (Ilya Repin's *An Unexpected Visitor*) with a number of different questions in mind. In a widely-reproduced figure, he depicted the series of fixations and saccades made in these different conditions, confirming that the viewing pattern in each case was very different. For example, when asked to give the ages of the people depicted in the scene, most fixations were on the faces of these characters. In contrast, when asked to "estimate the material circumstances of the family", many more fixations were made on the paintings, furniture and decoration in the room. Therefore the places where people looked were selected according to the information required by the task.

Yarbus' illustrations place a key emphasis on both where the observers were looking and the order in which they look there. Indeed, as discussed by Tatler et al., (2010) in their review of Yarbus' impact, he was particularly interested in the way

that eye movement scanning might be cyclical or idiosyncratic. When dividing up the viewing of an image into different time slices, he claimed that observers repeated a pattern of inspection. For example, when looking at a face, participants would iterate through the key features (eyes, mouth, nose) before starting again. Although different participants tended to look at similar locations, when the same person looked at an image on multiple occasions the resulting scanning patterns were even more alike.

In this section of the chapter, I will review methods and theory which examine these "scanpaths" or "scan patterns" in more detail. As we shall see, this goes beyond merely describing what is looked at, to quantifying the sequence of eye movements as a whole.

Scanpath theory links eye movements to a cognitive model for visual recognition

As we have seen, the viewing patterns of a number of people looking at an image may cluster on certain regions of detail. What determines the order or temporal structure of their fixations? Noton and Stark (1971) advanced the theory that the execution of a sequence of eye movements is intimately bound up in the processes of encoding and recognizing an image. Noton and Stark were influenced by Yarbus' observation of cyclical and repeating eye movements, and they used recordings from those inspecting simple line drawings. In particular, in each observer's pattern of inspection they claimed to see a "scanpath": a "fixed path, followed intermittently but repeatedly by a subject's eye while he views a pattern" (p. 933). This observation was based on judging the similarity within and between viewers, and in particular on recordings from the same person looking at the same image on two occasions. Noton

and Stark proposed that some elements of an individual's scanpath was repeated on subsequent viewings.

This qualitative observation was used to support a theory of human pattern recognition. According to "scanpath theory", when a visual pattern is encountered for the first time, its local features are stored in memory. Crucially, these sensory memory traces are combined with a representation of the eye movements made to view the pattern—a motor memory trace. The result is a sensorimotor network: a series of connections between nodes representing features and those representing the eye movements. When the pattern is encountered again at a later date, this network is reactivated, so that successive eye movements in the sequence are repeated, and the visual features are verified against the memory trace.

On the one hand, scanpath theory was an overly simplistic and impractical account of how people see and remember. It is not clear how features are integrated, or how such networks could scale up to all the different sorts of patterns that we can recognize, and all the situations in which we view them. On the other hand, it pre-dated much more modern accounts of embodied perception, such as Barsalou (1999), which see motor simulation as a crucial part of perception. For those interested in eye movements, the notion of scanpaths as generated top-down, based on individual memories and representations is a radically different approach to the research on visual saliency discussed in the first part of this chapter.

Methods for scanpath comparison

I began this chapter by discussing some methods for quantifying the eye movements from multiple observers looking at an image. In general, these methods

were based on the spatial position of fixations. For example, heatmaps tell us something about the spatial spread and consistency of fixations, and how they align with items in the scene. If we are interested in the *order* in which certain regions are fixated, one option is to evaluate measures of fixation timing (e.g., how early objects get fixated in different scenes or conditions, see Table 1 and Figure 2). However, this requires clear regions of interest and analyses only a small number of fixations (i.e., those which land on the regions). They are also generally aggregated across many trials and observers. If scanpaths really do feature predictable sequences of fixations, then these methods will probably not be able to detect them.

An alternative way of analyzing scanpaths is to compare them more holistically, over space and time. The central problem here is how to take two scanpaths (A and B) and compute their pairwise similarity. This problem has been addressed in detail by several authors (Privatera & Stark, 2000; Foulsham & Underwood, 2008; Dewhurst et al., 2012; Cristino et al., 2010; See Bischoff et al., this volume).

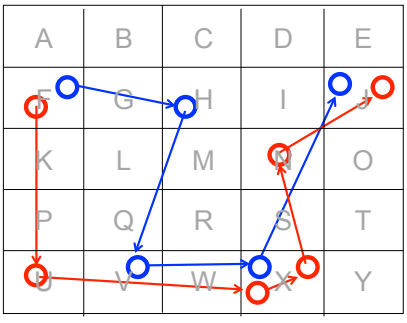
Perhaps the most straightforward approach would be to measure the linear distance between the fixations in each scanpath. Similar scanpaths will contain fixations which are close to each other in space. This is the approach that was taken by Mannan et al., (1995) when they studied fixations in pictures with various levels of image manipulation. But to which of the fixations in scanpath B should a given fixation in scanpath A be compared? There are a number of options (the closest, the average of all others, that which occurred at the same time), but all of these result in un-intuitive values when there are big differences in the distributions of the fixations in A and B. Moreover, measuring linear distance cannot easily take into account the

sequence of the fixations, meaning that two scanpaths which visit the same locations in a completely different order will seem highly similar.

Instead, a range of authors have proposed scanpath comparison methods based on aligning fixation sequences (see Textbox 3, and Bischoff et al., this volume). These have the advantage that they represent both spatial and sequential aspects of the scanpaths. However, they can be complex to calculate, and there is currently no clear consensus on which method to use. The preferred analysis will depend on whether there are clear regions of interest and the aim of the comparison. Whichever method is chosen, it is important to evaluate similarity values with care. Often, the aim of the analysis is to investigate whether scanpaths are more (or less) similar than some kind of chance or baseline expectation. Because of the global biases discussed earlier in the chapter, scanpaths will probably be similar, to a certain degree, even when they originate from different scenes and observers. A good approach, therefore, is to also compute some control comparisons, against which the experimental values can be evaluated.

Textbox 3: Sequential scanpath comparison

Several of the scanpath comparison algorithms that have been developed are based on the "string edit" or Levenshtein distance, which was applied to eye movements by researchers such as Brandt and Stark (1997). First, scanpaths are coded as sequences of characters based on regions of interest (Fig. A). Then, the similarity between these sequences is quantified by calculating the number of steps or edits required to transform one into the other.



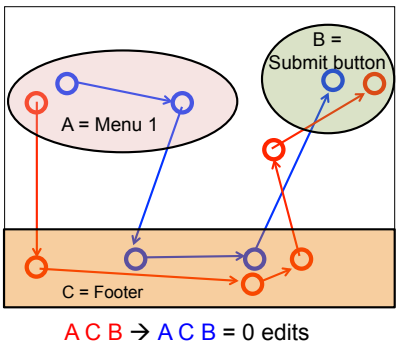


Fig. A. Two scanpaths are compared as strings, using either a grid or predetermined regions of interest.

Cristino et al., (2011) proposed a more sophisticated approach (ScanMatch) which takes into account fixation durations and spatial and non-spatial relationships between ROIs.

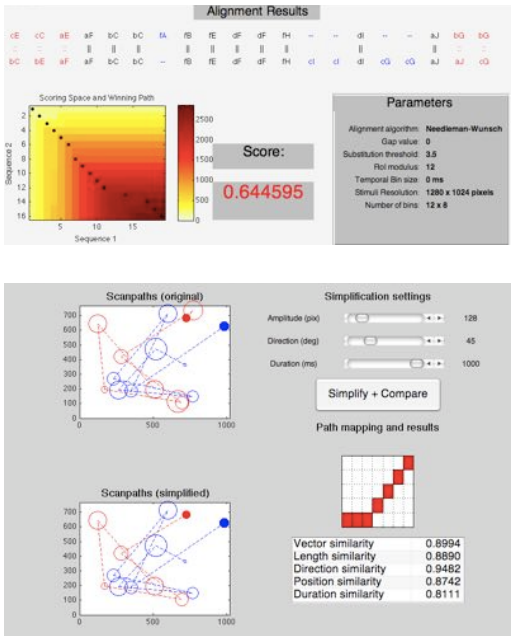


Fig. B. ScanMatch and MultiMatch toolboxes give similarity scores for a pair of scanpaths.

Dewhurst et al., (2012), described an alternative (MultiMatch), which aligns and compares scanpaths as simplified vectors. This allows scanpaths to be compared across multiple dimensions. Both ScanMatch and MultiMatch are available as MATLAB toolboxes (Fig. B).

Tests of scanpath theory confirm that participants are idiosyncratic in where they look

Now that we have considered how scanpaths can be represented and compared, we can return to the central predictions of scanpath theory: i) that participants will repeat eye movement sequences when recognizing a previously seen image; and ii) that this recapitulation will trigger successful recognition. Noton and Stark (1971) addressed i) by making qualitative judgements about the similarity of scanpaths at encoding and test. Subsequent work used transition matrices or the string edit distance to quantify similarity, confirming that there was some similarity between viewings (Stark & Ellis, 1981; Choi, Mosely & Stark, 1991). However, these experiments used rather simple stimuli and few observers, and it was not clear how the similarity measurements should be interpreted.

In Foulsham and Underwood (2008) and Foulsham et al., (2012), we returned to the question of scanpath similarity, using 45 complex natural scenes which were viewed by 21 observers as part of a memory test. Figure 6 shows an example of the resulting scanpaths. Using a variety of different comparison methods, we confirmed earlier reports that participants did show some similarity between viewings. This similarity was certainly not 100%, and because the image is the same, any differences must be due to memory and the demands of the task. However, scanpaths from the same person viewing the same image were more similar than two different people viewing the same image. In this sense, the eye movements of a particular individual seem to be idiosyncratic.

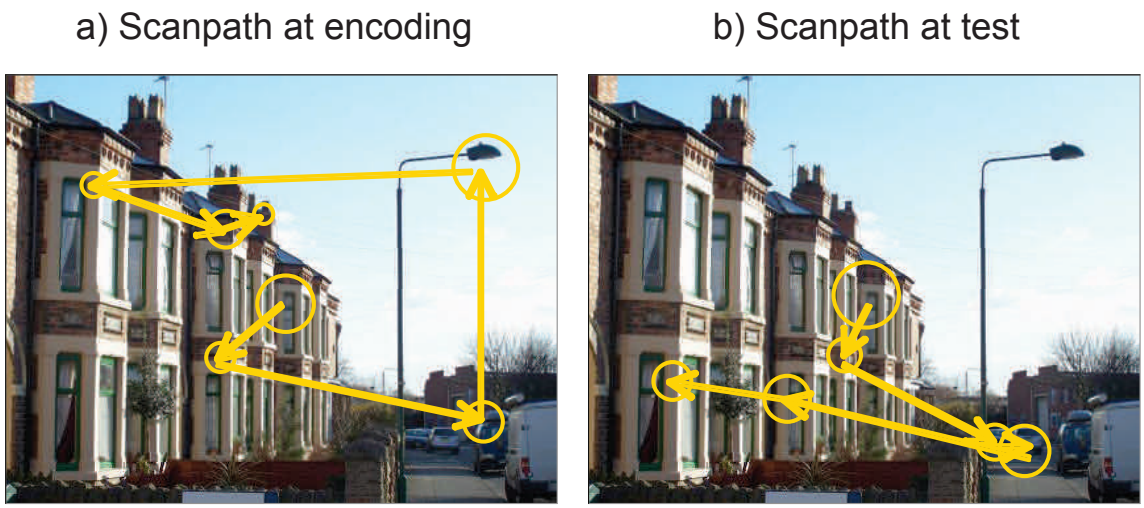


Figure 6. An example of two scanpaths made by the same person viewing the same scene, once while learning the image (a), and once when recognizing it later during a memory test (b). Figure redrawn from Foulsham et al. (2012). Note that in this figure fixation duration is represented by the size of the circles.

It is clear that the scanpaths made by the same person viewing an image on multiple occasions are not identical, and indeed, it is possible for people to recognize some images without making any eye movements. Therefore the strong version of scanpath theory can be rejected. Moreover, the modest similarity that does occur could be caused by a variety of factors: the re-presentation of salient items; a consistent reaction to scene elements; or idiosyncratic, systematic tendencies in eye movements which persist over time. So what of prediction ii) above, that eye movement recapitulation enhances recognition memory? The evidence that images with closely repeating scanpaths are recognized more accurately is inconclusive (Foulsham et al., 2012). It is also not clear whether repeating eye movements could be said to be causing memory, or the other way around. A much better test of the role of eye movements in image memory, therefore, is to manipulate scanpaths and observe the results.

Manipulating scanpaths can affect memory

Scanpaths in picture viewing arise when participants freely view an image, and thus they provide a measure of unconstrained, natural behavior. However, often a researcher may need to introduce constraints, so that information in the image is acquired in a certain order or central and peripheral material is controlled. Manipulating fixations in natural images can therefore be a useful tool. One possibility is to use a gaze-contingent procedure, which changes presentation based on voluntary eye movements. For example, Zelinsky and Loschky (2005) investigated memory in scenes by ending the trial when a certain number of objects had been fixated. Foulsham et al., (2013) used “moving windows” of different shapes to

encourage a different pattern of scanning, with vertical "slits" leading to more vertical eye movements.

Holm and Mantyla (2007) coerced participants viewing landscape paintings to follow a certain scanpath. Observers were required to fixate a sequence of dots, superimposed on the image. Later on in the experiment, looking back to the same locations was associated with explicit recognition. In Foulsham and Kingstone (2013b), we used a similar technique to probe the causal role of eye movements, providing a direct test of scanpath theory. In our first experiment, we manipulated the scanpath during the learning phase of the experiment, as had been done by Holm and Mantyla. Rather than freely viewing each image, participants saw a series of square patches, one at a time, which simulated a sequence of fixations. After viewing 48 scenes in this fashion, there was a memory test where the task was to indicate whether the displayed image had appeared in the first half of the experiment. During this "test" block, the full scenes were displayed and the question was whether unconstrained eye movements would follow the sequence from that image in the learning block. The results showed that indeed the imposed scanpath affected viewing at test, but only when the image was subsequently recognized correctly.

In subsequent experiments, we reversed the procedure, allowing free viewing during the learning phase but constraining the scanpath at test (Foulsham & Kingstone, 2013b, Experiments 2-5). This meant that the sequence of fixations at test could be manipulated to be either the same as those made by that observer when first seeing the image, or drawn from different locations. These experiments were technically challenging and required that the fixations be written to a data file during the learning phase, and then retrieved to constrain the test phase. If repeating a scanpath really activates a stored memory of the pattern, then we would expect

recognition of previously seen images to be better when we forced participants to look at the image in the same sequence as they had on the first encounter. This was the case when recognition was compared with a condition presenting random patches. However, when the control condition presented patches from the eye movements made by another individual, or the same fixation locations but in a scrambled order, there was no memory advantage. In other words, there is not anything especially useful about repeating your own scanpath, in its set order. Fixations select memorable locations, but in the case of complex scenes replaying a scanpath does not seem to have a clear causal affect on memory in the way proposed in scanpath theory.

Eye movements provide information about recognition and imagery

Scanpath theory seems to be unable to account for the particular relationship between eye movements and memory observed in Foulsham and Kingstone (2013b), because it places special importance on one's own scanning pattern, which we found does not necessarily enhance memory. Eye movements vary in multiple ways, and although some of this variability appears to be systematic within an observer, supporting a link with memory processes has proven difficult. However, scanpaths clearly are useful measures, both for characterising viewing patterns in a holistic fashion and for investigating memory. At the most basic level, eye movements and attention are gatekeepers to our sensory input, and so they must constrain what we end up seeing and storing in memory. Eye movements and memory continue to be explored in a variety of stimuli, particularly in face recognition (Althoff & Cohen, 1999; Schwedes & Wentura, 2012).

Scanpath theory and methods have also been applied to visual imagery, where there continues to be a debate about the extent to which eye movements are functional. In a typical experiment, participants view an image or other spatial pattern, and they then visualise or imagine this image later while looking at a blank screen. In these conditions, observers seem to spontaneously make eye movements, which may partially reflect items in the remembered image. Laeng and Teodorescu (2002) and Johansson et al., (2012) argued that preventing eye movements during retrieval led to poorer imagery, although it is not always clear whether the precise pattern of eye movements made is important for functional benefits.

7. Applications and implications

This chapter has described a range of methods for investigating eye movements in scenes through region-of-interest analyses, comparisons with saliency maps, and quantifying scanpaths. Because the focus has been on relatively complex stimuli which vary across many dimensions, these techniques are well suited to applying to a range of different contexts. Eyetracking has become a popular tool for researchers far beyond the traditional realm of cognitive psychology and vision, and here I will mention a few particularly active interdisciplinary topics.

Perception of art

Like Yarbus and Buswell, some of my examples have been drawn from experiments where participants viewed works of art (e.g., Box 1, which shows Gainsborough's painting *Mr and Mrs Andrews*). As with more theoretical work in scene perception, eyetracking experiments with art have addressed the consistency between observers, as well as how this may be affected by the techniques of the artist and viewer expertise.

Moving beyond the truism that people look at areas of detail in a painting, DiPaola, Riebe and Enns (2010) investigated the subtle techniques that Rembrandt and other artists use to guide the eyes. DiPaola et al. selectively modified the rendering of portraits to test the idea that artists enhance the detail of one side of the face in order to induce a particular viewing pattern. Sure enough, participants spent more time on a textured eye region than on the other side of the face, and this affected

judgements of the quality of the art. Thus artists may have evolved techniques which affect eye movements, and this may even be true in the colour balance of abstract artworks (Nodine, Locher & Krupinski, 1993).

Eyetracking has also been used to evaluate the viewer's experience of looking at a piece of art. Nodine et al., (1993) were among the first to investigate how artistic training might affect this experience as well as viewing patterns. The results suggested that untrained viewers focus more on individual objects, while experts were more sensitive to composition. This was reflected in more "specific" scanpaths targeted at relationships between objects. More recent work has shown that artistically-trained individuals spend more time on structural features (Vogt & Magnussen, 2007). Naïve observers may also change their gaze patterns when speaking about their interpretation of a painting, becoming more systematic (Klein et al., 2014). Thus eye movements continue to show promise in this interdisciplinary field. Indeed, eyetracking has even been incorporated into a large exhibition in an art gallery, resulting in data from thousands of visitors (Wooding, 2000).

Marketing and websites

There is now a large industry which seeks to evaluate and improve marketing materials by using techniques from visual attention research and eyetracking (see section D of this volume). In advertisements, experimental research has often looked at the way in which observers scan combined text and images. For example, Rayner et al. (2001) found that people normally first read the text associated with print advertisements, before looking at the product image. Wedel and Pieters (2008) provide a useful review of how eyetracking has been applied to marketing, as a now

dominant measure of the attention paid to branding. Although it is clear that there are things that marketers can do to increase the attention paid to their advertisements (e.g., by making the brand and advertisement larger), the impact on actual purchasing is less persuasive. Current research using eyetracking to investigate marketing has also investigated sequential scanpaths (Pieters, Rosbergen & Wedel, 1999) and the role of bottom-up saliency (Van der Lans, Pieters & Wedel, 2008). Moreover, the central fixation bias appears to have an effect on the products which are noticed and ultimately chosen (Atalay et al., 2012).

The visual exploration of websites has also provided a measure for marketers. For example, a large number of studies and findings are discussed by Nielsen and Pernice (2010). Among the most popular claims from this research are that participants show an "F-shaped" pattern when reading web content (distinguished by a large horizontal exploration which tapers off further down the page) and that observers frequently ignore advertisements ("banner blindness"). More relevant for the present chapter, both saliency and scanpaths have been investigated in the context of websites (Holmberg et al., 2014; Josephson & Holmes, 2002).

Computer vision

Computational models of visual saliency, benchmarked against human eye movement data, continue to be extremely popular in a range of computer vision applications. Borji and Itti (2013) review some of the many models that have been developed, along with their applications. Because it is assumed that fixation (and attention) is the first stage in cognitive processing, visual saliency may help build

computers which can recognize objects (Walther et al., 2005). Saliency can also be the first step in processes of image and video segmentation.

Interestingly, the focus on eye movements has both influenced and been influenced by developments in robotics. In particular, the acknowledgement that artificial visual systems benefitted from an active sensor (which moves like the eye), contributed to the complementary movements known as Active Vision (Ballard, 1991; Findlay & Gilchrist, 2003). Saliency has been implemented as a way for robots to cut down on their visual input and learn to move and localize themselves in space (e.g., Siagan & Itti, 2009).

8. Conclusions and limitations

In this chapter, I have discussed eyetracking results from people looking at pictures, and I have done so largely within the framework of saliency and scanpaths. Visual saliency provides an explicit and implemented model of going from simple visual features to predictions for complex scenes and images. However, it is clearly also a very limited approach for describing actual eye movement behavior. The weak empirical effects of saliency on fixations do, however, allow us to focus on ways of representing and manipulating task knowledge so that more explicit predictions can be made about where people will look in a given situation.

Eye movement data can be challenging to analyse. The research I have discussed shows that we should not forget the many ways that eye movements over pictures may vary, in terms of fixation position, saccade directions and so on. One often comes back to the observation that eye movements are very far from being uniformly distributed across a display, and this systematicity comes from both the stimulus and the observer. In particular, researchers should be mindful that some "ways of looking" may emerge more often, based purely on biases in the oculomotor system (the origins of which may be learned or biological). Although there does not appear to be a straightforward relationship between memory and scanpaths, there is a regular and idiosyncratic component to the way that we move our eyes.

There are of course limitations to the scope of what I have discussed, and to the conclusions which can be made from observing eye movements in scenes. There are many other descriptions of individual differences, emotional states and task instructions making a difference to where people look in images, and these fall under the general umbrella of top-down attention. It remains to be seen whether or not these

effects interact with visual saliency. Although most research on scene perception has been carried out with healthy, typical observers, complex images have also been used to investigate neuropsychological and developmental disorders (e.g., Freeth et al., 2011; Foulsham et al., 2009; Tseng et al., 2013).

Recording where people look in complex images can tell us a lot about attention and cognitive processing, but it cannot tell us everything. Fixation and attention are not synonymous. Covert attention can certainly be allocated separately from fixation, but there are few studies examining this in complex stimuli, and thus its role in naturalistic viewing is unknown. Importantly, we should also be cautious at using results from picture-viewing to make conclusions about attention in the real world. The vast majority of the research that I have described uses small, static scenes, presented without context in a highly constrained setting. In other words, it is far removed from the real world, which contains cues such as sound and motion and in which observers are immersed and free to move, where concepts such as a "central bias" may be meaningless. Nevertheless, those continuing to investigate looking at pictures will find much to discover.

9. Suggested readings

Henderson (2003) gives a good summary of the basic methods and findings in scene perception.

Foulsham (2015) reviews the state of the art regarding eye movements in scene perception, covering some of the same material as this chapter but with more detail and scope.

Itti and Koch (2001) provide a comprehensive introduction to computational modeling of attention, including a shorter description of their saliency map model.

Tatler et al., (2011) provide a detailed critique of saliency models and the picture viewing paradigm.

Le Meur and Baccino (2013) summarise some of the commonly used methods for comparing fixations and model predictions.

Holmqvist et al., (2013) is a detailed textbook on eye movement methodology, with particularly comprehensive coverage of scanpath comparison and other measures.

10. Questions for discussion

Why are there so many measures based on eye movements in scenes? Are they all necessary?

What is the best way of defining “saliency” in terms of eye movement control?

How useful is the Itti and Koch saliency map model for predicting human eye movements?

Is analysis of scanpaths necessary for investigating cognition and scene perception?

Are scanpaths random? If not, why?

What can experiments investigating eye movements in scene perception tell researchers working in marketing and other applied domains?

11. Bibliography

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 997–1010.

Atalay, A. S., Bodur, H. O., & Rasolofoarison, D. (2012). Shining in the center: central gaze cascade effect on product choice. *Journal of Consumer Research*, 39(4), 848–866.

Ballard, D. H. (1991). Animate vision. *Artificial intelligence*, *48*(1), 57-86.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(04), 637-660.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision research*, *49*(24), 2992-3000.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185-207.

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, *13*(10), 18.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience, 9*(1), 27-38.

Bruce, N., & Tsotsos, J. (2005). Saliency based on information maximization. In *Advances in Neural Information Processing Systems* (pp. 155-162).

Buswell, G. T. (1935). *How people look at pictures*. University of Chicago Press.

Choi, Y. S., Mosley, A. D., & Stark, L. W. (1995). String Editing Analysis of Human Visual Search. *Optometry & Vision Science*, *72*(7), 439-451.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41-51.

Cristino, F., Mathot, S., Theeuwes, J., & Gilchrist, I. (2010). ScanMatch: A novel method for comparing saccade sequences. *Behavior Research Methods, 42(3),* 692-700.

DiPaola, S., Riebe, C., & Enns, J. (2010). Rembrandt's textural agency: A shared perspective in visual art and science. Leonardo 43(2), 145-151.

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, *44*(4), 1079-1100.

Dickinson, C. A., & Intraub, H. (2009). Spatial asymmetries in viewing and remembering scenes: Consequences of an attentional bias?. *Attention, Perception, & Psychophysics*, *71*(6), 1251-1262.

Dodge, R., & Cline, T. S. (1901). The angle velocity of eye movements. *Psychological Review*, *8*(2), 145.

Ehinger, K. A., 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), 2.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 3.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford University Press.

Foulsham, T., Barton, J. J., Kingstone, A., Dewhurst, R., & Underwood, G. (2009). Fixation and saliency during search of natural scenes: the case of visual agnosia. *Neuropsychologia*, *47*(8), 1994-2003.

Foulsham, T., & Kingstone, A. (2013a). Optimal and preferred eye landing positions in objects and scenes. *The Quarterly Journal of Experimental Psychology*, *66*(9), 1707-1728.

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14-25.

Foulsham, T., & Kingstone, A. (2013b). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General*, *142*(1), 41.

Foulsham, T., Dewhurst, R., Nyström, M., Jarodzka, H., Johansson, R., Underwood, G., & Holmqvist, K. (2012). Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Journal of Eye Movement Research*, *5*(4), 3.

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception?. *Perception*, *36*, 1123-1138.

Foulsham, T., Kingstone, A., & Underwood, G. (2008). Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision Research*, *48*(17), 1777-1790.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 6.

Foulsham, T. (2015). Scene Perception. To appear in Fawcett, Risko & Kingstone (Eds.). *The Handbook of Attention*. MIT Press.

Freeth, M., Foulsham, T., & Chapman, P. (2011). The influence of visual saliency on fixation patterns in individuals with Autism Spectrum Disorders. *Neuropsychologia*, *49*(1), 156-160.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in Neural Information Processing Systems* (pp. 545-552).

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*, *7*(11), 498-504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. L. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. van Gompel, M. Fischer, W. Murray & R. W. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537-562). Amsterdam: Elsevier.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850-856.

Henderson, J. M., & Pierce, G. L. (2008). Eye movements during scene viewing: Evidence for mixed control of fixation durations. *Psychonomic Bulletin & Review*, *15*(3), 566-573.

Holmberg, N., Sandberg, H., & Holmqvist, K. (2014). Advert saliency distracts children's visual attention during task-oriented internet use. *Frontiers in Psychology*, *5*.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Holm, L., & Mäntylä, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, *35*(7), 1664-1674.

Itti, L., & Baldi, P. F. (2005). Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems* (pp. 547-554).

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

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10), 1489-1506.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194-203.

Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1289.

Josephson, S. & Holmes, M. E. (2002). Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *ETRA '02: Proceedings of the 2002 symposium on eye tracking research and applications*, ACM, New York, 43-49.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Proceedings of the 12th International Conference on Computer Vision* (pp. 2106-2113). IEEE.

Klein C, Betz J, Hirschbuehl M, Fuchs C, Schmiedtová B, et al. (2014) Describing Art – An Interdisciplinary Approach to the Effects of Speaking on Gaze Movements during the Beholding of Paintings. PLoS ONE 9(12): e102439.

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

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054-16059.

Laeng, B., & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science, 26*(2), 207–231.

Land, M. F. (1993). Eye-head coordination during driving. In *Systems, Man and Cybernetics, 1993.* (pp. 490-494). IEEE.

Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, *45*(1), 251-266.

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

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565.

Mannan, S., Ruddock, K., & Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9*(3), 363-386.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, *2*(11), 547-552.

Mackworth, N. H., & Thomas, E. L. (1962). Head-mounted eye-marker camera. *Journal of the Optical Society of America (1917-1983)*, *52*, 713.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8), 17-17.

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, *421*(6921), 370-373.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205-231.

Nielsen, J., & Pernice, K. (2010). *Eyetracking web usability*. New Riders.

Nodine, C. F., Locher, P. J., & Krupinski, E. A. (1993). The role of formal art training on perception and aesthetic judgment of art compositions. *Leonardo*, 219-227.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*(9), 929.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20-20.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: a computational model of fixation durations in scene viewing. *Psychological review*, *117*(2), 382.

Nyström, M., & Holmqvist, K. (2008). Semantic override of low-level features in image viewing-both initially and overall. *Journal of Eye-Movement Research*, *2*(2), 2-1.

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

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397-2416.

Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research*, 424-438.

Posner, M. I., Rafal, R. D., Choate, L. S., & Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, *2*(3), 211-228.

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

Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied*, *7*(3), 219.

Rothkopf, C. A., Ballard, D. H., & Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, *7*(14), 16.

Schwedes, C., & Wentura, D. (2012). The revealing glance: Eye gaze behavior to concealed information. *Memory & Cognition*, *40*(4), 642-651.

Siagian, C., & Itti, L. (2009). Biologically inspired mobile robot vision localization. *IEEE Transactions on Robotics*, *25*(4), 861-873.

Stark, L., & Ellis, S. R. (1981). Scanpath revisited: Cognitive models of direct active looking. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye Movements: Cognition and Visual Perception* (pp. 193- 226). Hillsdale, NJ: Erlbaum.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision research*, *45*(5), 643-659.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, *11*(5), 5.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, *17*(6-7), 1029-1054.

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *i-Perception*, *1*(1), 7.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97-136.

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Current opinion in neurobiology*, *13*(4), 428-432.

Tseng, P. H., Cameron, I. G., Pari, G., Reynolds, J. N., Munoz, D. P., & Itti, L. (2013). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, *260*(1), 275-284.

Underwood, G., & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59*(11), 1931-1949.

Unema, P. J., Pannasch, S., Joos, M., & Velichkovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, *12*(3), 473-494.

Van der Lans, R., Pieters, R., & Wedel, M. (2008a). Competitive Brand Salience. *Marketing Science*, 27(5), 922–931.

Vig, E., Dorr, M., & Cox, D. (2014). Large-scale optimization of hierarchical features for saliency prediction in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2798-2805).

Vogt, S., & Magnussen, S. (2007). Expertise in pictorial perception: eye-movement patterns and visual memory in artists and laymen. *Perception*, *36*(1), 91-100.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, *19*(9), 1395-1407.

Walther, D., Rutishauser, U., Koch, C., & Perona, P. (2005). Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Computer Vision and Image Understanding*, *100*(1), 41-63.

Wedel, M., & Pieters, R. (2008). A review of eye-tracking research in marketing. *Review of Marketing Research*, *4*(2008), 123-147.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202-238.

Wooding, D. S. (2002). Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 518-528.

Wurtz, R. H., & Goldberg, M. E. (1972). Activity of superior colliculus in behaving monkey. III. Cells discharging before eye movements. *Journal of Neurophysiology*, *35*(4), 575-586.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum press.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*(4), 787.

Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 154-164.

Zelinsky, G. J., & Loschky, L. C. (2005). Eye movements serialize memory for objects in scenes. *Perception & Psychophysics, 67*(4), 676–690.

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), 32-32.