# Sequential Selection, Saliency and Scanpaths

by

Ramin Fahimi

A thesis submitted to
The Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements
of the degree of

Master of Science

Department of Computer Science
The University of Manitoba
Winnipeg, Manitoba, Canada
Oct 2018

Thesis advisor
Author

**Dr. Neil Bruce**
**Ramin Fahimi**

**Sequential Selection, Saliency and Scanpaths**

# Abstract

Visual saliency and eye movements have been well studied, mostly in the capacity of predicting topographical spatial saliency maps. In this thesis, we examine the problem of sequential selection and sampling of image content in detail. Careful scrutiny is applied to existing metrics for measuring success of sequential selection strategies, and a new family of metrics is proposed with an intuitive interpretation and that provides more discriminative power in revealing differences between viewing patterns or computational models. This is accompanied by experimentation based on classic strategies for simulating sequential selection from traditional representations of saliency, and deep neural networks that produce sequences by construction. Experiments provide strong support for the necessity of sequential analysis of attention and a roadmap for moving forward.

# Contents

# List of Figures

# List of Tables

# Acknowledgments

I would like to begin by thanking my advisor Dr. Neil D.B Bruce. His thoughtful insight to solving problems has continuously helped me to find solutions presented in this thesis. He is an extremely smart and kind man who has always been very supportive of me. He has taught me how to have the spirit of a researcher and stay strong in ups and downs. Neil, you are incredibly talented and it has been an absolute pleasure to work with you.

I would also choose this moment to thank my committee members, Dr. Yang Wang and Dr. Jason Morrison, who have provided valuable suggestions and constructive feedback in perfecting my thesis. Thanks very much for your time.

I consider myself very lucky and honored to have such fine, brilliant people in the lab leading me through the accomplishment of this thesis. I wish you all success. I have an special thanks for Ms. Shiva Shabro for pushing me through this thesis. You are amazing.

My parents and grandparents has always filled my hearth with immense love and support. I love you guys so much. You have moved mountains for me to be here and follow my passion. I'll always try to make you proud.

In the end, I want to recognize University of Manitoba, department of computer science and Neil for the the financial support. I am truly thankful and It wouldn't have been possible without your support.

*This thesis is dedicated to my parents.*

# Chapter 1

# Introduction

## 1.1 General Introduction

Human and animal brains have significant yet limited computational power, especially when considering the amount of sensory data available to them. Given this limitation, the brain appears to have evolved to provide an efficient solution to actively control the stream of information. Attention is the process of selectively concentrating on some portion of available information, at the expense of ignoring other perceivable parts. It is a combination of behavioral and cognitive processes and it may be defined subjectively or objectively. It has been widely researched in cognitive and perceptual psychology and regardless of discipline, the usage has been identified to be critical to Information Reduction and filtering.

Visual attention is a form of attention that allows humans and animals to control and allocate their limited visual processing resources. The form of control varies across species. Attention is a complex process that may also include physical actions like eye or head movements. In humans, it has been posited that more than 60% of the brain is involved in processing visual data; a high portion of activated capacity. It is not that there is too much information, the problem is that each component of each stimulus can be matched to many different objects or types of scenes in our memory resulting in a explosion of potential interpretations. In general, attention is a crucial mechanism for animals and humans survival.

Sequences of successive eye movements during scene exploration are described as visual scanpaths. The human eye reveals a rapid drop-off in visual acuity and most neurons are positioned in front of the fovea as can be seen in Fig 2.1. This means that humans have to move their gaze to effectively place a region of interest (ROI) in front of their fovea where they have the highest visual acuity. During sequences of eye movements, the eyes tend to pause briefly on certain locations that are known as fixations. Fixations are followed by saccades, which involve rapid changes in eye position driven by a ballistic action that rapidly accelerates the rotation of the eyes to land at a new target (and with some imprecision). Extracting visual information takes place during the fixation period and the whole process enables humans to do fine-grained analysis of the scene. A more detailed review of different types of eye movements is discussed in section 2.1.1. From one side, not all of the spatial locations will be viewed and from the other side, the selective process and ordering becomes important. Other parameters are also involved in this. Visual inhibition of return (IOR) facilitates foraging by discouraging re-examination of recently fixated locations [66]. Due to the limited capacity of visual working memory (VWM), effects of IOR gradually fade over time, which allows the possibility to re-fixate previously attended regions.

Saliency and visual attention have been studied in a computational context for decades. Much of the emphasis of studies has been focused on the notion of saliency, defined by Koch et al. [43] as a two-dimensional topographic representation of conspicuity for every location (pixel) in the image. This has been used to characterize the spatial density of visual fixations [13], objects of interest in a scene [35; 36], the degree of objectness [48] of pixels or bounded regions and other less common definitions. Other efforts have also sought to attach task to this, examining these densities from a top-down perspective [2]. Moreover, the notion of attention has become increasingly prevalent in the context of deep-learning wherein recurrent mechanisms or reinforcement learning drive the gradual solution to a problem, in some instances involving spatial selection [76].

While this problem has been studied in some detail, the overwhelming majority of work in this domain considers a static problem wherein the distribution of gaze, attention or salience is marked by its spatial distribution. However, spatial selection by an attentive mechanism is inherently a sequential sampling process in humans and many artificial vision systems. This evidently gains additional importance as sequential attention mechanisms in artificial vision systems are becoming increasingly prevalent. There may be a strong bias to examine certain parts of a scene right away before moving on to other interesting regions. On the other hand, there may be several equally interesting regions of a scene and no particular order. This is a characteristic that most modeling and analysis to date has failed to address in detail. While some efforts have been devoted to analyzing sequential models of fixation, the case may be made (as it is in this thesis) that existing metrics fall short in adequately capturing model performance, and similarities and differences among sequences of selected regions.

## 1.2 Motivation and Contributions

Established algorithms for saliency prediction have been improving over the years in predicting patterns derived from gaze data. However, in the vast majority of cases, they have been tested with datasets containing few salient objects and more importantly, the focus of this analysis and measurement of success has been based on 2D topographical representations of what is salient. Moreover, with more crowded and occluded scenes including many natural images, the attentive behaviour of humans varies significantly from one person to another. Given that this is not about right or wrong patterns of viewing, sequential analysis could be a more insightful avenue for investigation.

In general, simulating and modeling the human visual system has the advantage that it results in human-like behavior. This is beneficial for systems that need to interact with humans in a natural manner. Seeking simple solutions to derive a sequence of fixation points while exploring a scene could also help attentive models produce higher accuracy for their specific task. High-level strategies for optimal information gathering are key to the field of computer vision. Most applications in computer vision function primarily in a passive way. Meanwhile, active control over the acquisition of image data is fundamental to efficient development of robust and general computer vision solutions for unconstrained environments. The advent of Deep Learning has been something of a revolution in machine vision capabilities and has brought a significant boost in the capabilities and performance of solutions. However, even in this domain there is no escaping the large amount of computing power necessary to adequately treat input especially if the desire is a solution that is possible to run in real-time. Incorporating attention in models could help converge to a better solution. This also raises the question of what value can be gained, or what can be understood in comparing the attentive policy of models built on machine learning models for specific tasks with humans.

Motivated by the aforementioned observations, we seek to revisit the space of metrics currently used in the domain of analyzing gaze trajectories, viewing patterns and saliency in order to arrive at a consensus on intuitive interpretations of inter-sequence distances, and also towards redefining metrics that produce meaningful and significant contrast among observations. In the balance of this thesis, we present metrics and experimentation that advocates for alternative metrics to any appearing in the literature deemed ScanPath Plausibility (SPP). This is tested in considering a wide variety of extant saliency models coupled with a selection mechanism.

Contributions of this thesis are therefore as follows:

- A detailed analysis and review of metrics available for scanpath comparison. Before this, metrics were discussed from a theoretical standpoint without empirical experiments or quantification of metric behavior.

- Revealing strengths and weaknesses of metrics from both an axiomatic and empirical standpoint along with recommendations for analysis.

- Constructing the first benchmark of computational saccadic models based on a series of simple post-processing techniques on saliency models, and the very limited set of models that produce sequences by construction.

- Comparative studies based on information in a sequence and showing the wealth of information comparing to static representation.

## 1.3 Thesis Outline

The remainder of this thesis is organized as follows: In Chapter 2, background and related work are reviewed and discussed. Chapter 2 also introduces a variety of metrics covered in this thesis including static and sequential metrics for comparing gaze patterns, and their relationship to inter-observer congruency. Chapter 3 demonstrates an empirical study on the behaviour of sequential metrics and provides an in depth benchmark of the performance of saliency algorithms while discussing various factors related to benchmarking and performance. Furthermore, this chapter introduces the ScanpPath Plausibility metric (SPP) for measuring Inter-observer congruency also providing an upper bound for models performance. Chapter 4 delves into some of the cases where sequential analysis can be more enlightening than static measures. In particular, it is revealed that interpretation of data changes when viewed through the lens of sequential metrics, which underscores their value and necessity. Finally, Chapter 5 concludes the thesis describing limitations and directions for future research.

## 1.4 Open data & Source Code

The Python implementation of all experiments can be accessed through the following link: https://github.com/rAm1n/msc-thesis. The evaluation data including the materials, datasets, individual scanpaths, metrics and figures can be downloaded from the same link.

# Chapter 2

# Background and Related Works

In this section, related works and studies similar to those considered in this thesis are reviewed. Starting with human gaze behaviour, different types of eye movements are discussed. Following this two most common methods of computational modeling of eye tracking data will be reviewed: i. Static saliency map prediction or ii. temporal saccadic models. In the literature, the extent of agreement across observers is measured and quantified according to an inter-observer congruency (IOC) score. This plays a role primarily in benchmarking in measuring the capacity for the raw human fixation data to predict behavior of other humans in the experiments, and to provide a bound on how well an algorithm may perform in the best case. As we use IOC extensively in Chapter 3, this is defined in more detail.

## 2.1 Visual Attention

Attention is the process of selectively concentrating on some facet of available visual information, while ignoring other perceivable parts. Visual attention allows us to focus our visual processing resources, tuning the sensitivity of neurons to relevant stimuli, and filtering out distractors and noise in a dynamic process. A more detailed discussion of the factors involved in this process while taking into account neuro-biological considerations is necessary for the purposes of the understanding the data presented in this thesis.

Two kinds of visual attention exist. The first involves eye movements (overt orienting) while the other shifts within the focus of the mind without moving the eyes (covert orienting). Helmholtz [57] first demonstrated that covert fixations or shifts of attention over a scene without eye movements were real. With that said, we are far from having a complete neuro-biological understanding despite now knowing much more about eye movements, components of the human imaging system and overt attention. In general, three main factors play a role in visual attention systems: Eye movements, the foveal gradient of resolution and also neural processing. Eye movements are rich source of information. As the fovea captures information in the highest detail, given the high density of cone receptive cells in fovea, the eye moves around quickly to areas containing certain stimuli so that light from regions of interest falls directly on the fovea. Studies strongly suggest that an early pre-attentive stage in which eye movements are purely stimulus driven helps in the creation of a mental model of a scene. Following this, an attentive stage of fixation helps to form a model that accounts for a supposed goal. During this process, perception of the scene is fabricated through continuous analysis by the brain of the time-varing image captured on the retina [12].

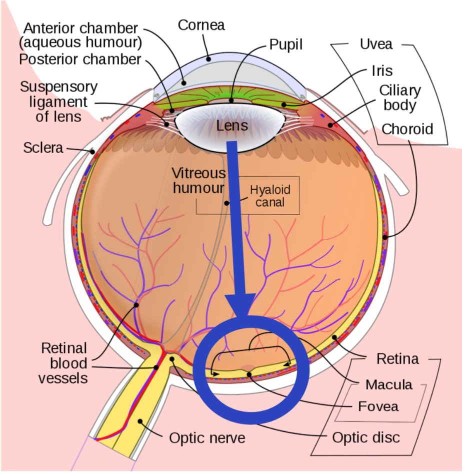


Figure 2.1: Inside humans eyes - Fovea [1]

### 2.1.1 Eye movements

There has always been a debate among researchers as to whether eye movements are stimulus driven or task/experience/memory driven. Yarbus study showed that the task at hand can significantly impact eye-movement behaviour, however, an attracting stimuli leads to a much more consistent behaviour [77]. In general, from a top-down point of view, eye movement patterns are related to personal characteristic. It can either be conscious as in performing tasks at hand or unconscious driven by factors such as health, age, gender or personality. From a bottom-up approach, the process is related to the visual stimulus; It might be low-level based on local image features or high-level (e.g. based on social context). The last factor is related to oculomotor system, such as the spatial bias that comes from the tendency to fixate the center region of an image or scene.

Eye movements that take place while freely viewing a scene provide significant information about context or an observers intent. Eye-movements are the most obvious external manifestation of a change of visual attention and one might consider more specifically different types of eye movements in this context. We are principally concerned with a sequence of fixations or pause following by saccades.

Saccades are rapid changes in eye position, allowing the eyes to jump from pointing at one position in space to another. The human visual system has a rapid drop-off in acuity and most photo-sensitive neurons are positioned right in front of the lens, at the fovea. Therefore, fixations aim to bring objects of interest into the fovea. The process of extracting visual information takes place during the fixation period. Planning and execution of a single saccade takes about 150 to 200ms and the average span of a saccade is 20-40ms. Mathematically they are normally defined by amplitude, velocity, direction, duration and latency.

There are 6 category of eye-movements when broadly considering their definition based on physiological parameters. Vestibular-Ocular Reflex are non-voluntarily reflex functions to stabilize images on the retinae during head movement (or external movement of the world) by producing eye movements in the direction opposite to head movement. The objective of these movements is to retain the point of interest on the fovea during head and body movements. Smooth Pursuit of the eyes happens when voluntarily tracking moving stimuli. The difference between smooth pursuit and saccadic movements is that saccades happen as uneven jumps. Nystagmus or

| | |
|---|---|
| Saccade | voluntary fast & ballistic movements |
| Vestibular-Ocular Reflex | compensatory changes in eye position as head moves |
| Smooth Pursuit | voluntary tracking of moving stimuli |
| Nystagmus | Reset by a primitive saccade similar to reaching the orbit limit |
| Optokinetic Nystagmus | low-frequency rotations at constant velocity. |
| Vergence | Coordinated movements - converging or diverging |
| Torsion | Coordinated rotation, dependent on head tilt and eye elevation. |

Table 2.1: Summary of eye-movements

dancing eyes and Optokinetic Nystagmus, are a condition of involuntary eye movement, acquired in infancy or later in life, that may result in reduced or limited vision. Nystagmus may be caused by congenital disorders and they can be detected where the eye aims to remain fixed but move rapidly instead. Vergence is the simultaneous movement (rotation) of both eyes in opposite directions to obtain or maintain single binocular vision. This is useful in cases where depth changes. The eyes may move in opposite directions inward to examine a closer object, or outward to observe an object farther away. Finally, Torsion is the movement that brings the top of the eye toward the nose (intorsion) or away from the nose (extorsion) [73].

### 2.1.2 Eye Tracking Datasets

As eye trackers have become cheaper and affordable to a larger audience, more datasets have been published. At the time of writing this thesis, the list of datasets available according to well established MIT benchmark of saliency [17] includes 26 datasets and continues to increase. In order to delve into various aspects of sequential analysis for eye-tracking data, different datasets have been employed, a list of which can be found in table 2.2.

| Name | Size | Observer | duration | Fixations | Type |
|---|---|---|---|---|---|
| OSIE [75] | 700 | 15 | 3 | ✓ | 👁 |
| CROWD [40] | 500 | 16 | 5 | ✓ | 👁 |
| PASCAL-S [52] | 850 | 12 | 2 | ✓ | 👁 |
| CAT2000 [10] | 2000 | 18 | 5 | ✓ | 👁 |
| FIGRIM [16] | 2787 | 15 | 2 | ✓ | 👁 |
| LOWRES [75] | 168+25 | 15 | 3 | ✓ | 👁 |
| SALICON [38] | 10000 | 60 | - | ✓ | 🖱 |

Table 2.2: Available Datasets for training & evaluating models of gaze behaviour. Note that this list derives from careful consideration of all available datasets, their curation, and modifications in how information is extracted to engage in both spatial and temporal analysis.

## 2.2 Computational modeling of eye movements

Data collected from eye-tracking is a very complex signal that can carry different meanings and allow for different types of analysis. Eye movements may be different in nature and in their impetus based on a variety of aspects including but not limited to age, health, experience, task, attractiveness of stimuli and many others. This data is a rich (albeit noisy) reflection of the inner workings of the human brain and has been used to understand the behaviour and the under-laying cognitive process leading to perception. It is now much cheaper to collect large sets of eye-tracking data and with

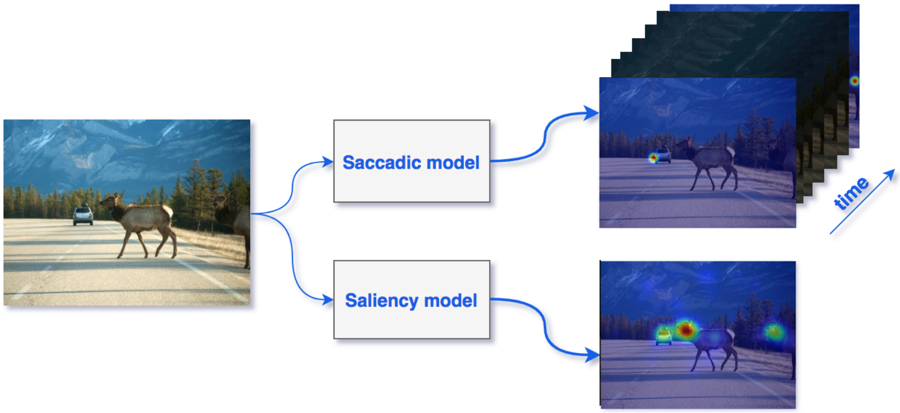


Figure 2.2: Saliency (static) models vs. Saccadic models

ethical concerns limited mostly to nonscientific experiments, eye tracking is one of the most effective and efficient solutions to gather rich data that allows for understanding many aspects of brain behaviour. At a high level, looking at the data recorded by eye-trackers reveals sequence of somewhat stochastic fixations that have four main dimensions **Position, Shape, Duration** and **Order**. There has been a longstanding effort trying to replicate and model any of these four fundamental features by different computational strategies. Two major categories of studies have either focused on modeling position and areas of attention, or tried to replicate plausible scanpaths and eye-movements covering all four dimensions. In this section, techniques used in both categories are reviewed in detail and the following chapters leverage some of these considerations in focusing on evaluating results from computational models.

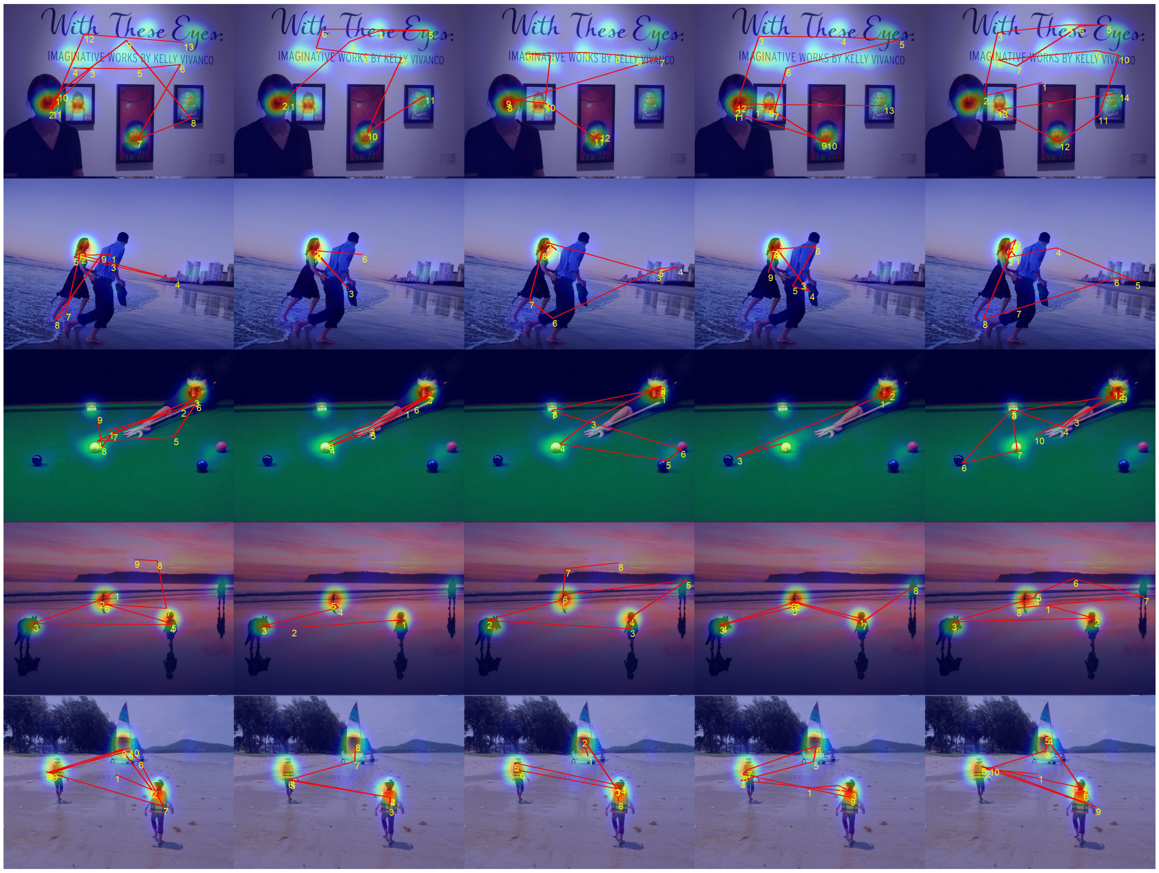


Figure 2.3: Scanpaths and salient regions – the ordering reveals a compelling difference among observers, even when over loci of fixations are consistent.

### 2.2.1 Spatial Distribution of Fixation Positions

The most extensively studied technique in attention prediction is based on computing a saliency map (heat-map) which represents the amount of attention received by each pixel (spatial information). The salient locations in a scene depend on the composition of the scene including areas of contrast, objects present and prior experience or knowledge. The emphasis of these models has largely been based on non-ordered prediction of fixations and variety of solutions have been proposed. Saliency maps model bottom-up attention in free viewing tasks without any particular goal.

A series of images is presented to different observers without providing any particular instructions about what to search for in the images. This type of strategy is assumed by most models of image saliency. Such analysis reveals image understanding related to color, shape, orientation, or motion irrespective of specific objects. The post-processed data comes from blurring the fixation map of all observers and Koehler [44] showed that saliency maps show better approximates explicit judgments than gaze points in fixation data.

Traditional models have adopted different techniques including but not limited to information theory [14; 68], cognitive approaches [43; 37; 58] graphical models [33] and Bayesian analysis [72; 78]. Learning based approaches, specifically Deep Learning has raised the bar on performance for many problems in computer vision, including saliency prediction. In the last few years, multiple deep learning based models [47; 70; 46; 39] have been proposed and achieve promising results for the static problem. Reinforcement learning [79] and Adversarial learning [61] are other types of learning based solutions that have revealed some success. Evaluation metrics are discussed in detail in the next section.

### 2.2.2 Scanpath Analysis and Saccadic Models

There exists another category of work exist that aims to add temporal information to the previous models. These approaches aim to predict scanpaths or the sequence of fixation points. This task in nature is relatively harder than predicting heat-maps because of the disagreements of different viewers (i.e, different viewers might visit same salient objects but with different order). Early interest in saccadic sequences was influenced by Noton and Starks scanpath theory [60], which was largely devoted to the memory encoding and recall of images. The theory suggests that individuals recognizing a previously seen image follow a scanpath similar to their initial viewing of the image. Meaning that observers reproduce the same sequences of eye movements when recognizing or imagining a stimulus. Noton and Stark limited their study to visual inspection to determine whether two scanpaths were similar but there have since been efforts [29; 27; 15] trying to examine their findings based on scanpath similarity metrics. This has been proven to be correct by Foulsham et al. [29] and they further [27] found that also scanpaths are more similar within an individual than between individuals. Even if visual memory is not heavily influenced by scanpaths, there are nevertheless a number of applications that can still make good use of fixation modeling. This includes all commercial applications of saliency analysis including graphical figures [31], active camera control for robotics tasks [55][65], omnidirectional camera systems [30], self-driving vehicles [45] etc. Bottom-up saliency models have been used to estimate the scanpath. Itti and Kochs models [37; 43] utilize a Winner Take All (WTA) selection for generating scanpaths using the predicted saliency map. Le Meur [56] includes a bottom-up generated saliency map, oculomotor biases, and IOR in their proposed framework to predict scanpaths.

Deep learning solutions have also been proposed for saliency driven scanpath generation [7]. Assens et al. [7] proposed a feed forward Convolutional Neural Network (CNN) to predict saliency maps quantized in time. Instead of predicting one density map, they predict 10 maps which represent saliency according to the chronological period of time. However, the generated scanpath accuracy is highly dependent on the quality of saliency volume and sampling strategies. Ngo [59] proposed an encoder-decoder framework that extracts image features using a CNN and then forwards those features to standard Long Short Term Memory (LSTM) to generate scanpath points at different time steps. Similarly, Chen et al. [19] proposed to use a ConvLSTM as the decoder to generate scanpaths while learning Inhibition of Return (IOR) along with bottom-up attention. PathGAN is a recently proposed solution based on adversarial learning. A detailed analysis of metrics for Scanpaths are discussed in 3 including associated analysis.

## 2.3 Eye Movement Analysis - Evaluation

Given the right question, eye movements can provide deep insight into the inner workings of the mind. Sometimes the goal is to analyze and understand the underlying mechanisms and sometimes replicating visual attention is the goal. Looking at the evaluation part in general, when the goal is to understand similarity measures among gaze patterns are critical. Proper similarity metrics for comparing human eye movements can used as a starting point for more complex analyses. Looking for patterns and matching with external signals while also finding common sub-sequences of eye movements is another common approach in evaluation criteria. As outlined in the previous section, in this section evaluation criteria for both cases will be discussed.

### 2.3.1 Spatial Approaches

Spatial approaches for analyzing eye-tracking data, especially those that only consider position, have a long standing history in the literature. Spatial metrics are highly

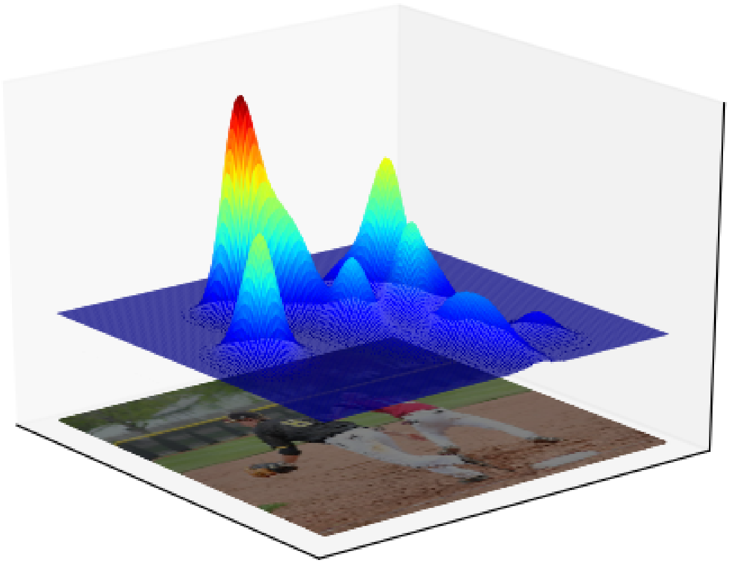
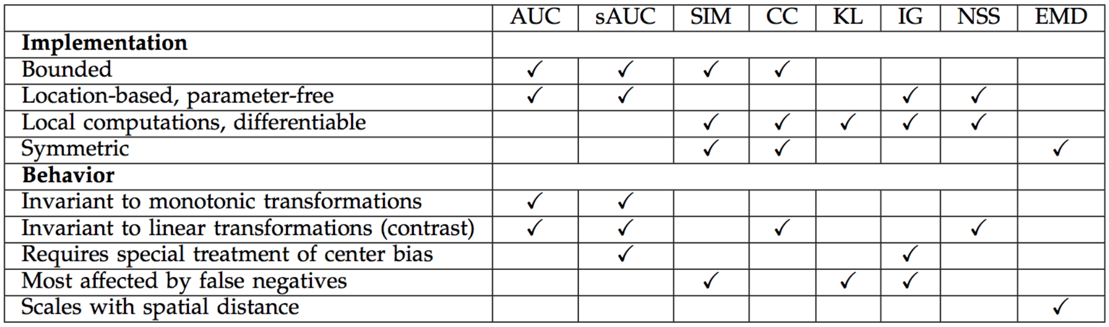


Figure 2.4: 3D visualization of a saliency map. The height of the surface represents the expected degree of interest in a particular image location.

trusted and have been used in the vast majority of research efforts surrounding eye-tracking. The notion of "Saliency" has changed meaning and normally refers to the topological maps or spatial distributions of attention usually visualized as a heat-map overlayed on an image. They can be grouped in two categories: location-based and distribution-based. These metrics sometimes show inconsistency in how they rank saliency models and can often leave performance open to interpretation. There are Empirical and deep investigations are already published in the literature on analyzing different aspects of metrics [26; 51; 49; 18] and much has been said specifically about spatial metrics. Table- 2.3.1 published by Bylinskii et al. [18] summarizes important aspects of these metrics.



| | AUC | sAUC | SIM | CC | KL | IG | NSS | EMD |
|---|---|---|---|---|---|---|---|---|
| **Implementation** | | | | | | | | |
| Bounded | ✓ | ✓ | ✓ | ✓ | | | | |
| Location-based, parameter-free | ✓ | ✓ | | | | ✓ | ✓ | |
| Local computations, differentiable | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Symmetric | | | ✓ | ✓ | | | | ✓ |
| **Behavior** | | | | | | | | |
| Invariant to monotonic transformations | ✓ | ✓ | | | | | | |
| Invariant to linear transformations (contrast) | ✓ | ✓ | | ✓ | | | ✓ | |
| Requires special treatment of center bias | | ✓ | | | | ✓ | | |
| Most affected by false negatives | | | ✓ | | ✓ | ✓ | | |
| Scales with spatial distance | | | | | | | | ✓ |

Table 2.3: A summary of static metrics for comparing gaze patterns (by Bylinskii et al. [18])

### 2.3.2 Scanpath Analysis Approaches

Different metrics have been proposed to compare two scanpaths. The simplest form of these metrics are methods based on constructing a grid and use string-edit distances [20; 23]. More sophisticated vector based methods try to geometrically align scanpaths (e.g. MultiMatch [22]). These methods do not necessarily provide a measure of models of visual exploration that is readily interpretable and moreover, these methods often rely heavily on free parameters (e.g. the grid resolution). Table 2.4 highlights the most commonly used metrics.

In the following section, metrics used in this thesis will be reviewed in detail. Given two scanpaths in two dimensional Cartesian coordinates, namely $P = (p_1, p_2, ..., p_n)$, $Q = (q_1, q_2, ..., q_m)$ with scanpath lengths N and M on stimuli with sides W and H, this discussion examines metrics that have been introduced for comparing sequences based on various criterion. In this chapter, the goal is to thoroughly investigate

| | Metric | abrv | Quantization | Target |
|---|---|---|---|---|
| 1 | Euclidean distance | EUC | Direct | Position |
| 2 | Mannan distance [53] | MAN | Direct | Position |
| 3 | Eyeanalysis [54] | EYE | Direct | Position |
| 4 | Levenshtein distance [63] | LEV | Grid | Position, Order |
| 5 | ScanMatch [21] | SMT | Grid/Temporal | Position,Order,Duration |
| 6 | Hausdorff distance [34] | HAU | Direct | Position |
| 7 | Frechet distance [24] | FRE | Direct | Position, Order |
| 8 | Dynamic time warp [8] | DTW | Direct | Position, Order |
| 9 | Time delay embedding [74] | TDE | Direct | Position |
| 10 | MultiMatch Shape [22] | MM_S | Direct | Shape |
| 11 | MultiMatch Direction(angualr)[22] | MM_A | Direct | Direction |
| 12 | MultiMatch Length [22] | MM_L | Direct | Length |
| 13 | MultiMatch Position [22] | MM_P | Direct | Position |
| 14 | MultiMatch Duration [22] | MM_D | Direct | Duration |
| 15 | Recurrence [5] | REC | Radius | Position |
| 16 | Determinism [5] | DET | Radius | Fixation Trajectories |
| 17 | Laminarity [5] | LAM | Radius | Fixation Persistence |
| 18 | Corm [5] | COR | Radius | Leading/Following |

Table 2.4: Common metrics for evaluation of Scanpaths.

the effectiveness and focus of each metric, understand associated behaviours and finding similarities among them. An overview of the metrics included in this thesis is presented in Table 2.4. Shape, ordinal ordering, duration and other factors are critical to the extent to which metrics lend themselves to interpretation. These are explored systematically in Chapter 3.

### Euclidean distance

Euclidean distance or straight-line distance is one of the basic and initial metrics that was used in comparing scanpaths. This metric only works with scanpaths that have equal lengths. Assuming $l$ is the minimum length among two scanpaths, the distance can be calculated as the sum of the distances between each fixation pair:

$$D_{EUC}(P,Q) = \sum_{i=1}^{\min(N,M)} \sqrt{(X_P - X_Q)^2 + (Y_P - Y_Q)^2}$$

Euclidean distance has been internally used in many of the metrics. Given the simplicity and only supporting scanpaths with equal lengths, we won't include it our analysis.

### Levenshtein similarity (Edit distance)

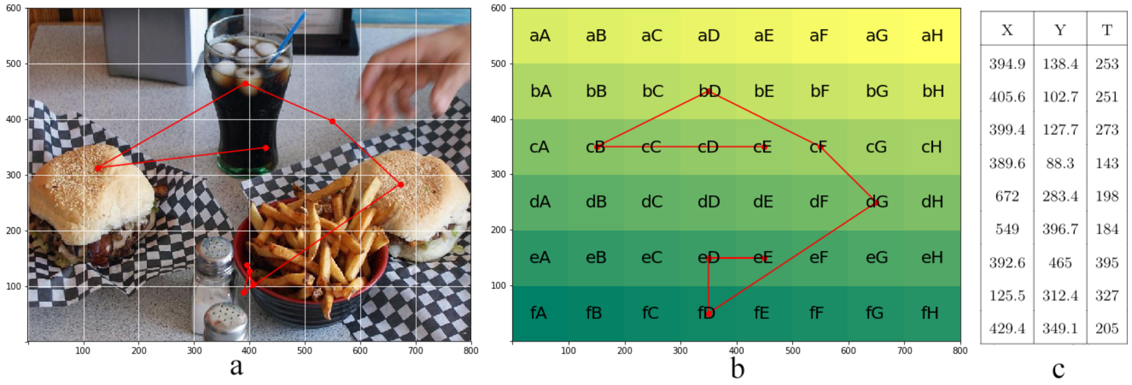
The use of Levenshtein similarity or edit distance [63] for the purpose of comparing scanpaths dates back to Noton and Stark [60] as one of the very first metrics used in literature. The Levenshtein algorithm originally compares two set of DNA strings $A(a_1, a_2, , ..a_a)$, $B(b_1, b_2, ..b_b)$ with lengths a and b based on the minimum number of insertions, deletions and substitutions to produce 2 identical sequences:

$$d_{A,B}(i,j) = \begin{cases} max(a,b) & if min(i,j) = 0 \\ min \begin{cases} d_{A,B}(i-1,j)+1 \\ d_{A,B}(i,j-1)+1 \\ d_{A,B}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & otherwise \end{cases}$$

To be able to use this algorithm for scanpath comparison, fixation coordinates needs to be converted to a string. As originally used in Noton studies [60], a static grid is overlayed on the stimulus to discretize areas of interest; Each grid cell is assigned an alphabetic character. Following this, a series of fixations can be represented using a string of characters and the Levenshtein algorithm can be used to compare strings representing them. As powerful as it may seem in considering ordinal information, there are various aspects related to this algorithm and other string-based metrics that aren't fully taken into account. The spatial position of fixations are not well represented and this can have consequences for the similarity metric. Two fixations that are spatially close, can land on different bins of the grid which could lead to an inaccurate distance. Similarly, fixations occupying the same bins are often grouped and represented with one letter. This has the effect of ignoring local differences in eye movement control and ongoing cognitive process. In general, the Levenshtein algorithm is blind to the semantic contents of an image and can't distinguish differences in substitutions of fixations from different regions. Moreover, duration of fixations is completely ignored in this algorithm.

### ScanMatch

ScanMatch can be seen as advanced and improved version of Levenshtein distance. It has recently emerged in the literature [21] and has been widely used in studies [62; 69; 28]. ScanMatch tries to solve some of the shortcomings of Levenshtein distance in considering semantic information and duration of fixations. Levenshtein distance treats all differences between strings equally. The edit cost of substituting, inserting and deleting is the same in all cases, and this results in being blind about information lying within other areas of interest. The advantage of this metric comes from use of NeedlemanWunsch pair-wise algorithm that aligns two string with one another to maximize a similarity score. The NeedlemanWunsch algorithm uses a substitution matrix containing all of the pairings in two strings. This makes it possible to weight the substitution operation for various areas of interest. For instance two pairings can have similar low-level information like color and luminance, or fall on the same semantically meaningful objects based on the goal of comparison when constructing the matrix.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| aA | aB | aC | aD | aE | aF | aG | aH |
| bA | bB | bC | bD | bE | bF | bG | bH |
| cA | cB | cC | cD | cE | cF | cG | cH |
| dA | dB | dC | dD | dE | dF | dG | dH |
| eA | eB | eC | eD | eE | eF | eG | eH |
| fA | fB | fC | fD | fE | fF | fG | fH |

| X | Y | T |
|---|---|---|
| 394.9 | 138.4 | 253 |
| 405.6 | 102.7 | 251 |
| 399.4 | 127.7 | 273 |
| 389.6 | 88.3 | 143 |
| 672 | 283.4 | 198 |
| 549 | 396.7 | 184 |
| 392.6 | 465 | 395 |
| 125.5 | 312.4 | 327 |
| 429.4 | 349.1 | 205 |

Figure 2.5: a) Visualization of scanpath record in c. b) pre-processing stage for transforming scanpaths to strings. ”eDeEeDfDdGcFbDcBc” without considering duration and” eDeDeEeEeDeDfDdGcFbDbDbDcBcBcBcEcE” when T=100ms.

ScanMatch metrics follow a similar pre-processing steps to Levenshtein distance for converting scanpaths to a sequence of strings. Following this, the objective is to fill a comparison matrix representing two strings on different axis based on the substitution matrix. The algorithms search for an optimal path through this matrix from top-left to bottom-right to calculate the similarity score and then the normalized score subject to the length of the sequence provides the final resulting score ranging between 0 and 1. A score means of identical scanpaths according to the pre-processing step and also most notably the substitution matrix. It’s important that substitution matrix should properly capture the connection between semantic or visual content among different areas of interest (AOI) but this shouldn’t be subjective. While ScanMatch has the advantage of not being blind to the semantic content, similar to Levenshtein distance, it still suffers from problems arising as the result of discretizing the AOIs and ignoring the exact positions of fixations. Also, ScanMatch doesn't respect similarities in shape. Two scanpaths can be very similar in terms of shape but have different scales. Similarly, a scanpath rotated in angular space won't be similar according to the metric.

One interesting feature that can be included in this process is the duration of fixations points. The duration of a fixation can be quantized according to a fixed temporal bin and for as long as the fixation stays, the string representing the AOI can be repeated in the overall sequence. For instance, if a fixation takes $t$ ms and the bin is T ms, the character representing the bin will be repeated for $\lceil \frac{t}{T} \rceil$ times.

### Mannan distance

Mannan distance [53] is a direct measure that considers fixations exact position rather than quantizing AOIs. For each fixation, the pair-wise euclidean distance with all fixations in other sequences is recorded in $d$. The Mannan distance is equal to the mean of minimum distances for each and every fixation when compared with all of other fixations in the other sequences.

$$D_{MAN} = [1 - \frac{D}{D_r}] * 100$$

$$D^2 = \frac{N * \sum_{j=1}^{M} \min d_{i,j}^2 + M * \sum_{i=1}^{N} \min d_{i,j}^2}{2 * N * M * (W^2 + H^2)}$$

Considering absolute positions in this way has both benefits and disadvantages. It doesn't need to quantize the space; it's simple and easy to calculate and also might make it possible to use raw data from eye-trackers. However, as each fixation is associated with its nearest neighbor, the ordinal information is mostly ignored and it softens the impact of individual fixations. This makes it very difficult to retain and consider one of the main characteristics of scanpaths, the nature of sequential information.

### Eyeanalysis

A few alternative approaches [29; 32; 71] have been proposed to rectify some of the shortcomings of Mannan distance. Eyeanalysis [54] proposes double mapping instead of mapping one fixation to one other fixation. This provides the lowest overall position variability but still may lead to several points in one scanpath being mapped to just one in the other. There exists also the possibility of including the duration of fixation. It's not a direct measure for ordering of fixations but would allow the temporal characteristics of the sequence to be taken into account to a limited degree.

$$D_{EYE} = \frac{(\sum_{i=1}^{M} \min d_{i,j}^2 + \sum_{j=1}^{N} \min d_{i,j}^2)}{max(N+M)}$$

### Dynamic Time Warping

Dynamic time warping [8] can compare two time-series with varying (and differing) lengths. Given two sequences with length N and M, DTW tries to build an NxM matrix in which each cell keeps the euclidean distance between the fixations with respect to their position in the sequence. Following this, it tries to find an optimal path to match both sequences while preserving three conditions: Boundary, Continuity and Monotonicity to make sure that the path respects the time. The optimal solution is the minimum among all the paths in order to get from (0,0) to (N,M).

**Algorithm 1** Dynamic Time Warp

```
procedure DTW($(P, Q)$)
    $D \leftarrow array[0..n, 0..m]$
    $D[:, 0] \leftarrow infinity$
    $D[0, :] \leftarrow infinity$
    $D[0, 0] \leftarrow 0$
    for $i \leftarrow 1 : N$ do
        for $j \leftarrow 1 : M$ do
            $D[i, j] \leftarrow minimum(D[i-1, j],\ D[i, j-1],\ D[i-1, j-1])$
            $D[i, j] \leftarrow D[i, j] + d(P[i], Q[j])$
        end for
    end for
    $returnD[n, m]$
end procedure
```

### Discrete Frechet distance

The Frechet distance[6; 25] is a measure of similarity between two curves. Unlike Hausdorff that doesn't respect the ordering of fixations, Frechet accounts for the location and ordering of points along the curves. Consider a dog walking along one curve and the owner walking along the other curve connected by a leash. With both walking continuously from their respective starting points, and varying their speed but not backtracking, Frechet distance between the two curves is the length of the shortest leash sufficient for traversing the curves:

$$D_{FRE(A,B)} = \inf_{\alpha,\beta} \max_{t\in[0,1]} \left\{ d\Big( P(\alpha(t)),\, Q(\beta(t)) \Big) \right\} \tag{2.1}$$

### Hausdorff distance

Hausdorff distance[34] represents the degree of mismatch between two sets by measuring the farthest distance from one set to the other. Hausdorff is mostly used for measuring the distance between two curves in space and is very useful for computer vision tasks. However, it doesn't consider temporal information. Given two Scanpaths P and Q, Hausdorff distance is defined as follows:

$$D_{HAU}(P,Q) = max(h(P,Q), h(Q,P)) \tag{2.2}$$

$$h(A,B) = \max_{a\in P} \min_{b\in Q} ||a-b||$$

In simple words, it measures how far is the maximum distance among the mapping between sets based on minimum distance.

### Time delay embedding

Time delay embedding[67] has been widely used in the study of dynamical systems and was introduced for scanpath comparison by Wang[74] et al. The last few metrics discussed in this section have meaningful and similar intuition behind them with slight differences, somehow involving euclidean distance. TDE tries to make consecutive sub-samples. P and Q go through the same pre-processing step which break each of them into consecutive samples with length K. The final output comes from calculating mean or maximum (Hausdorff) of pair-wise euclidean distance between sub-samples from each scanpath. The parameter K, plays a key role as change in ordinal order is one of 8 common categories of noises for scanpath data, which is discussed in depth in chapter 3. With a proper K, this metric is very powerful in handling this type of noise.

### MultiMatch

MultiMatch [56] is one the widely used metrics in the literature. It calculates the distance based on five components: Shape, length, direction, position and duration, fundamental characteristics of saccades between fixations. The main advantage of the MultiMatch method is that it provides several measures to choose from for assessing scanpath similarity, and each measure on its own captures a unique component of scanpath similarity. Given the multiplicity of measures, it remains difficult to assess at measure, or what set of measures, is most applicable in a given scenario. Furthermore, because each scanpath is initially simplified it is also not clear how robust each measure is to scanpath variations [3].

The algorithm follows a pre-processing stage where scanpaths are simplified in order to cluster fixations and saccades that are close together or represent local scanning. It starts with direction and continues with length and for as long as there are consecutive saccades that are smaller than a fixed threshold, they are either merged or removed. Some precautionary measures need to be taken into account when simplifying scanpaths. Merging saccades with small changes in direction while they are large in terms of amplitude might cause a loss of some important information. Following the pre-processing step, scanpaths are aligned based on their shape using Dijkstra algorithm. Alignment could happen based on any of other 4 parameters but it's common to do so based on shape. This alignment reduces the comparisons sensitivity to small temporal or spatio-temporal variations, and allows the algorithm to find the best possible match between the pair of scanpaths. Now, for each of the 5 parameters the difference is as follows:

MultiMatch Vector similarity is computed as the vector difference between aligned saccade pairs. the Length similarity is computed as the absolute difference in the amplitude of aligned saccade vectors and the position similarity is computed as the Euclidean distances between aligned fixations. All three of these measurements are normalized by the screen diagonal and averaged over scanpaths. The Direction Similarity is computed as the angular difference between aligned saccades, normalized by $\pi$ and averaged over scanpaths. Finally, the duration similarity is computed as the absolute difference in fixation durations of aligned fixations, normalized by the maximum duration and averaged over scanpaths.

### Recurrence, Determinism, Laminarity, Corm

Recurrence quantification analysis (RQA) has been successfully used for describing dynamic systems that are too complex to be characterized adequately by standard methods in time series analysis. It has been recently introduced for scanpath analysis by Anderson et al. [4]. Despite the simplicity of definition, this family of metrics has a clear interpretation and intuitive meaning which makes them able to capture global and local temporal characteristics of a sequence by a small number of RQA measures.

All four metrics are somehow based on recurrence matrix $r$. Two fixations are considered to be recurrent if they are closer than a fixed threshold $\rho$. The Closeness" function $d$ can be defined in several ways; It can be Euclidean distance or use a grid similar to Levenshtein or ScanMatch and fixations that fall on the same AOIs will be considered close. The recurrence matrix is presumed to be diagonal. If the length of scanpaths are not identical, one of them will be truncated to be in the same length (N).

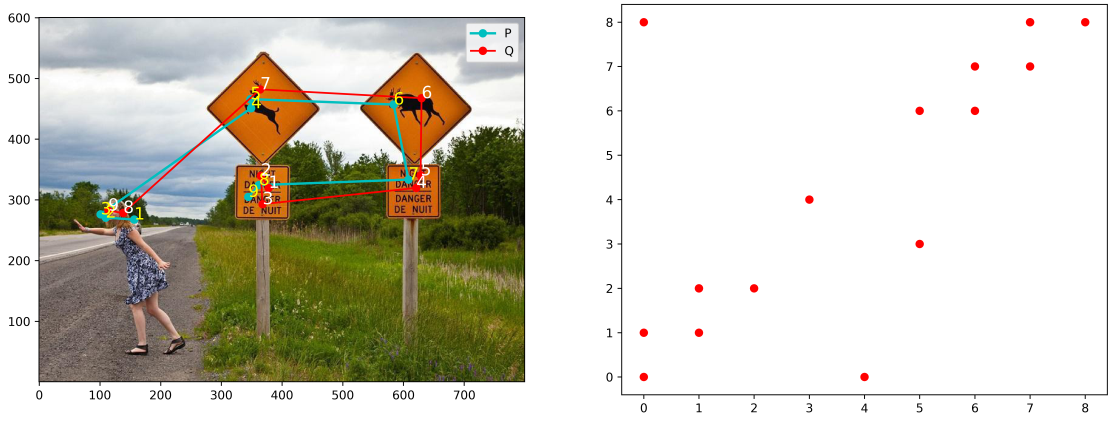


Figure 2.6: Recurrence matrix based on euclidean distance with $\rho = 2 * 24$(visual angle)

Let C be the sum of recurrences, i.e., $R = \sum_{i=1}^{N} \sum_{j=1}^{N} r_{ij}$ . Further, let $D_L$ be the set of diagonal, $H_L$ the set of horizontal, and $V_L$ the set of vertical lines in the cross-recurrence matrix, all with a length of at least L, and let $|.|$ denote cardinality. Therefore:

The ***Cross-Recurrence (REC)*** measures the percentage of fixations that match as are close between the two fixation sequences.

$$D_{REC}(P, Q) = 100 * \frac{C}{N^2}$$

The ***Determinism (DET)*** measures the percentage of cross-recurrent points that form diagonal lines and represents the percentage of fixation trajectories common to both fixation sequences. That is, it quantifies the overlap of a specific sequence of fixations, preserving the sequential information. The minimum line length of diagonal line elements was set to L = 2. Similarly, the ***Laminarity (LAM)*** represents locations that were fixated in detail in one of the fixation sequences, but only fixated briefly in the other fixation sequence(L = 2). Finally, the ***Center of recurrence mass (CORM)*** is defined as the distance of the center of gravity from the main diagonal, normalized such that the maximum possible value is 100. The CORM measure indicates the dominant lag of cross recurrences. Small CORM values indicate that the same fixations in both fixation sequences tend to occur close in time, whereas large CORM values indicate that cross-recurrences tend to occur with either a large positive or negative lag. [3]

$$D_{COR}(P,Q) = 100 * \frac{\sum_{i=1}^{\min(M,N)} \sum_{j=1}^{\min(M,N)} ((j-i)r_{ij}}{(\min(M,N)-1)C}$$

$$D_{DET}(P,Q) = 100 * \frac{|D_L|}{R}$$

$$D_{LAM}(P,Q) = 100 * \frac{|H_L|+|V_L|}{2*C}$$

### 2.3.3 Inter-observer Congruency

Eye-tracking data is a subjective source of information. As discussed in section-2.1.1 various aspects and prior experiences in human life can play a role in driving the movements of our eyes. Different people may find different locations of an image interesting and there can be little agreement between observers. This is one source of confusion for models trying to learn to predict saliency and but also makes it difficult to interpret evaluation scores and rankings among models. To address this issue, efforts have been made towards trying to find an upper bound on performance for predicting saliency across datasets, models or in particular, observers. In the literature, the extent of the agreement across observers has been introduced and quantified according to an inter-observer congruency (IOC) score.

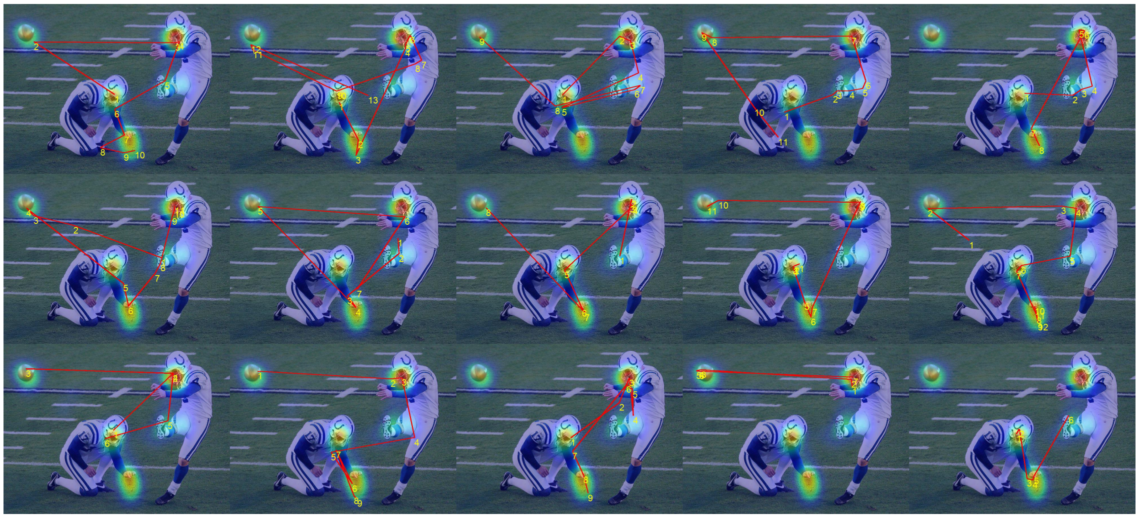


Figure 2.7: Visualization of 15 observers viewing the same image revealing differences in viewing

IOC score is also known as leave one out way approach. Given N observers, it uses (N-1) observer's fixations to predict the one left out then averages over the results of (N) evaluation scores per stimulus sample. Following Le Meur et al. [50] that tried to predict the IOC, other researchers also continued analyzing various aspects of this [11; 64; 42]. In [50], they showed that the correlation is not significant within a low confidence interval between predicted results and actual IOC scores. It was also noted that duration of viewing plays an important rule on the variability and free viewing data is less predictable than viewing data derived from specific tasks. Given the fact that ordering and time spent on each fixation is known in sequential analysis, we seek to revisit IOC scores for sequential data and inspect various dimensions of it including the sequential analogue of IOC.

## 2.4 Summary

Given the right question, eye movements can provide deep insight into the inner workings of the mind and can have many more applications in other areas including image comparison, object tracking, depth estimation and other areas. Broadly looking at the data recorded by eye-trackers, this is marked by a sequence of fixations involving a somewhat stochastic process that have four principal dimensions namely **Position, Shape, Duration** and **Order**. Spatial approaches for analyzing eye-tracking data only consider the position. They are highly established in the literature, trusted by many researchers and have a long history of use in gaze based analysis. As much as analysis has focused on the static perspective, relatively little has happened for sequential analysis; this is especially the case when comparisons are made between computational models. In this chapter, prior work on computational modeling of eye-tracking data, and metrics for evaluating the models have been reviewed. Moreover, the subtext surrounding this presentation also offers insight beyond the original work into the relative strengths and weaknesses of different approaches, and some general characteristics that may be important to consider in what makes a metric suitable. This exercise has also involved a comprehensive examination of datasets, selection of those that lend themselves to analysis of sequential viewing patterns, and scripts and methods to extract the relevant information from extant datasets that have been used almost exclusively for position based analysis. In the Chapter that follows, we make use of these various datasets to measure and study interpretability, strengths and weakness of sequential metrics.

# Chapter 3

# Sequential Analysis of Eye Movements

## 3.1 Introduction

### 3.1.1 What Can a Similarity Measure Do?

Scanpath similarity metrics can be used as a starting point for many more complex analyses. Collecting large amounts of eye-tracking data is easy and is getting easier, but analyzing the data in an appropriate way with the increasing wealth of information is not. It's important to develop new and powerful tools for the analysis of eye movement data. Similarity metrics can be used as a starting point for more complex analyses and developing and testing models. When it comes to similarity measures in a free-viewing task, two major approaches have normally been taken that are for model evaluation data analysis for different purposes. As discussed in Chapter

2, Saccadic models exists that aim to replicate human scanpaths. Humans exhibit complex behaviours when viewing an image, which all factor into observed position, order, shape and duration of fixations and saccades. As alluded to in Chapter 2, this can relate to personal characteristics, or alternatively, depend on the properties of stimuli (e.g. what's in the image). A similarity metric should be able to weigh differences and similarities among models, and in particular to differentiate and rank the models. From a different perspective, a good similarity measure should help answer a specific but common class of questions. As noted by Cristino et al. [54], "it allows you to cluster similar eye movement sequences together, or detect differences between predefined groups of eye movement sequences". They considered 4 common scenarios for experiments:

1. Detecting differences between predefined sets: Given sets of eye-tracking data of recorded subject with different conditions, the similarity measure should help to determine the effect of manipulating conditions of experiment. To answer this question, normally the average similarity of pairs of eye movement sequences within sets (defined by an experimental condition) should be larger than the similarity of sequence pairs between sets.

2. Supporting scanpath theory: A similarity measure should also evaluate whether two sequences of viewing that belong to the same person (encoding and recognition) viewing the same scene (effect of person) are on average more similar than two sequences of the same person viewing different scenes (effect of image) and two sequences of different people viewing the same scene.

3. Diagnostic use: Given known groups of data, a suitably designed similarity

measure (or metric) are able to assign exemplars from a new set of eye-tracking data to known categories based on similarities. For instance, eye-tracking has been used to diagnose Alzheimer's [9] and works reasonably well as a diagnostic tool.

4. Data-driven clustering: With a large set of data, a metric should differentiate well enough so that clustering techniques should be able to find similar clusters in data. This might comprise different people with similar viewing strategies, the same person viewing different content, different individuals viewing a common image or other explicit or implicit groupings.

### 3.1.2 What Makes a Metric Suitable?

It is difficult to assess the value of metrics without an associated reference frame. Some studies have produced models that reveal no significant difference when compared with human data. On the basis of this, there is the temptation to assume that the model is successfully replicating human gaze patterns. However, this can be equally the success of a model, or failure of a metric. Moreover, success at discriminating between different models according to performance may be strong but lack specificity. One example of this is analysis that is purely spatial (e.g. static AUC). Another example is metrics that consider only summary statistics of gaze patterns like MultiMatch.

Given this context, we can state some guiding principles (axioms) that should inform the choice of a suitable metric. Specifically the metric should:

1. Be capable of measuring distances between sequences of fixations.

2. Have an intuitive interpretation.

3. Be able to effectively capture the order, position, duration of fixations and shape of a scanpath[22].

4. Provide a level of sensitivity that allows for reasonable separation of models that produce good vs. poor sequences

5. Should not consist only of coarse grained saccade or fixation statistics

Each of these can be considered both in the context of the 4 principal measurements defined in Chapter 2, and in light of the guiding principles proposed by Cristino et al. [54].

### 3.1.3 What is Missing in Scanpath Comparison?

While the problem of saliency prediction has been studied in some detail, the overwhelming majority of work in this domain considers a static problem wherein the distribution of gaze, attention or salience is marked by its spatial distribution. However, spatial selection by an attentive mechanism is inherently a sequential sampling process in humans and many artificial vision systems. This evidently gains additional importance as sequential attention mechanisms in artificial vision systems are becoming increasingly prevalent.

There maybe be a strong bias to examine certain parts of a scene right away before moving on to other interesting regions. Alternatively, there may be several equally interesting regions of a scene and no particular order. This is a characteristic that most modeling and analysis to date has failed to address in detail. While some efforts have been devoted to analyzing sequential models of fixation, the case may be made that existing metrics fall short in adequately capturing model performance, and similarities and differences among sequences of selected regions.

Table- 2.4 shows 18 metrics for comparing scanpaths. This clearly illustrates the amount of interest in the subject but strangely shows the lack of exploration at the meta-level in the value of metrics for analyzing behaviours and outlining strengths and weakness of these metrics. In many cases, studies that used eye-tracking data, each introduced their own metric for analyzing the data (perhaps in support of showing a desired result) but little is known about the interpretation of their metrics.

Motivated by the aforementioned observations, we revisit the space of metrics currently used in this domain to arrive at a consensus on intuitive interpretations of inter-sequence distances, and also towards redefining metrics that produce meaningful and significant contrast among observations. The balance of this chapter, presents experiments that ultimately advocate for an alternative metric to any appearing in the literature, ScanPath Plausibility (SPP). While this is a novel metric, the essence of SPP builds on existing metrics albeit with some additional intuition related to the guiding principles established above. This is tested in considering a wide variety of extant saliency models coupled with a selection mechanism.

## 3.2 How do Metrics Measure Up?

As alluded to earlier in this thesis, a wide variety of different metrics have been proposed, and yet it is unclear the extent to which these are suitable for measuring similarity among sequences of gaze points. While many of these are grounded in measuring distances between trajectories or in computational geometry, there is a relative lack of domain knowledge inserted into the choice and use of such metrics. We therefore consider an alternative approach, and present careful analysis of information carried by different metrics and means of determining which are most discriminative. We have used various datasets for the analysis but most of the numbers are based on OSIE [75]; however, we have confirmed similar results on CROWD [40] and CAT2000 [10] datasets.

### 3.2.1 Imposters and Noise Handling

In order to better ground the interpretation of metrics, we have conducted experiments that consider human vs. human scanpath similarity. That is, given $n$ observers, one can perform a leave-one-out analysis, treating the instance left out as our exemplar, and using the remaining observers to predict gazed at points of the individual left out and determine a bound on expected model performance. We have carried out this experiment for the set of metrics considered measuring human vs. human similarity, and also human vs. imposter, where the imposter scanpath is a real scanpath but drawn from a different image. In this manner, we have the capacity to consider the extent to which scanpaths that are plausible *in general* conform to scanpaths that are plausible *for a particular image*. In practice, a good metric should be capable of discerning what sequences belong to a common image vs. those that are not specific to the image considered. In considering the overall distributions of scanpath distances $A$ and $B$ with mean and standard deviations $\mu_A$ and $\sigma_A$, and $\mu_B$ and $\sigma_B$ respectively, where $A$ is all instances of leave-one-out comparisons from the same

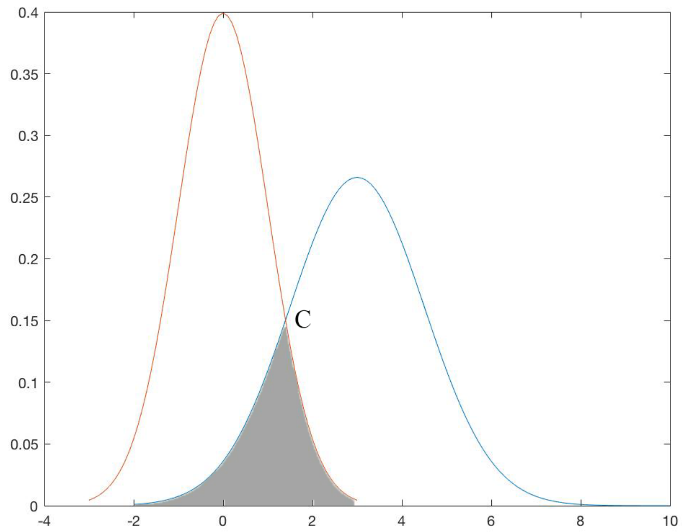


Figure 3.1: Intersection of two normalized distribution.

image, and $B$ is all instances of leave-one-out where the comparator is an imposter, we can readily calculate the overlap among the 2 distributions. This corresponds to the proportion of instances that belong to one distribution, but would be classified as belonging to the alternative. A lower degree of overlap is diagnostic of a more powerful metric for discerning these 2 groups (following classic decision theory). If C is presumed as the center of the intersection and F is the cumulative distribution function, the ratio of intersection is given by:

$$P(X_1 > c) + P(X_2 < c) = 1 - F_1(c) + F_2(c)$$

$$= 1 - \tfrac{1}{2}\mathrm{erf}\left(\tfrac{c-\mu_1}{\sqrt{2}\sigma_1}\right) + \tfrac{1}{2}\mathrm{erf}\left(\tfrac{c-\mu_2}{\sqrt{2}\sigma_2}\right)$$

where C can be calculated by:

$$c = \frac{\mu_2\sigma_1^2 - \sigma_2\left(\mu_1\sigma_2 + \sigma_1\sqrt{(\mu_1-\mu_2)^2 + 2\left(\sigma_1^2 - \sigma_2^2\right)\log\left(\frac{\sigma_1}{\sigma_2}\right)}\right)}{\sigma_1^2 - \sigma_2^2}$$

This assumes that histograms must follow a normal distribution. As shown in Figure- 3.2, we tried a normality test based on Agostino and Pearsons tests. The test combines skew and kurtosis to produce an omnibus test of normality and all the metrics rejected the null hypothesis. Final Results are shown in Table 3.2. It is immediately evident that a few metrics have poor power of discrimination while others stand out. It is also the case that some that lack discriminative power fail to meet some of the considerations defined under section- 3.1.2.

In addition to revealing behaviour of metrics the results of this experiment for SPP are unequivocal. The SPP distances show significantly higher capacity for separating these classes in part due to resilience to domain relevant noise that is intrinsic to inter-observer variability, but also given the high intrinsic variability of scanpaths that implies a large distance when an imposter is considered.

To take this analysis further, we ask a different question. Suppose inter-observer scanpath similarity is considered for a given image, and this is compared with inter-observer scanpath similarity where an increasing number of imposters is introduced to the observer pool. In this case, leave one out analysis is performed on the group for all N observers with K imposters (K < N-1). One can also consider how the distributions diverge as the observer pool is increasingly polluted with random scanpaths from other images. The results of this analysis appear in Figure 3.3. One can see that the SPP based on a min distance across observers is much less sensitive to noisy samples being introduced. Moreover when the sample is entirely or almost entirely imposters, the

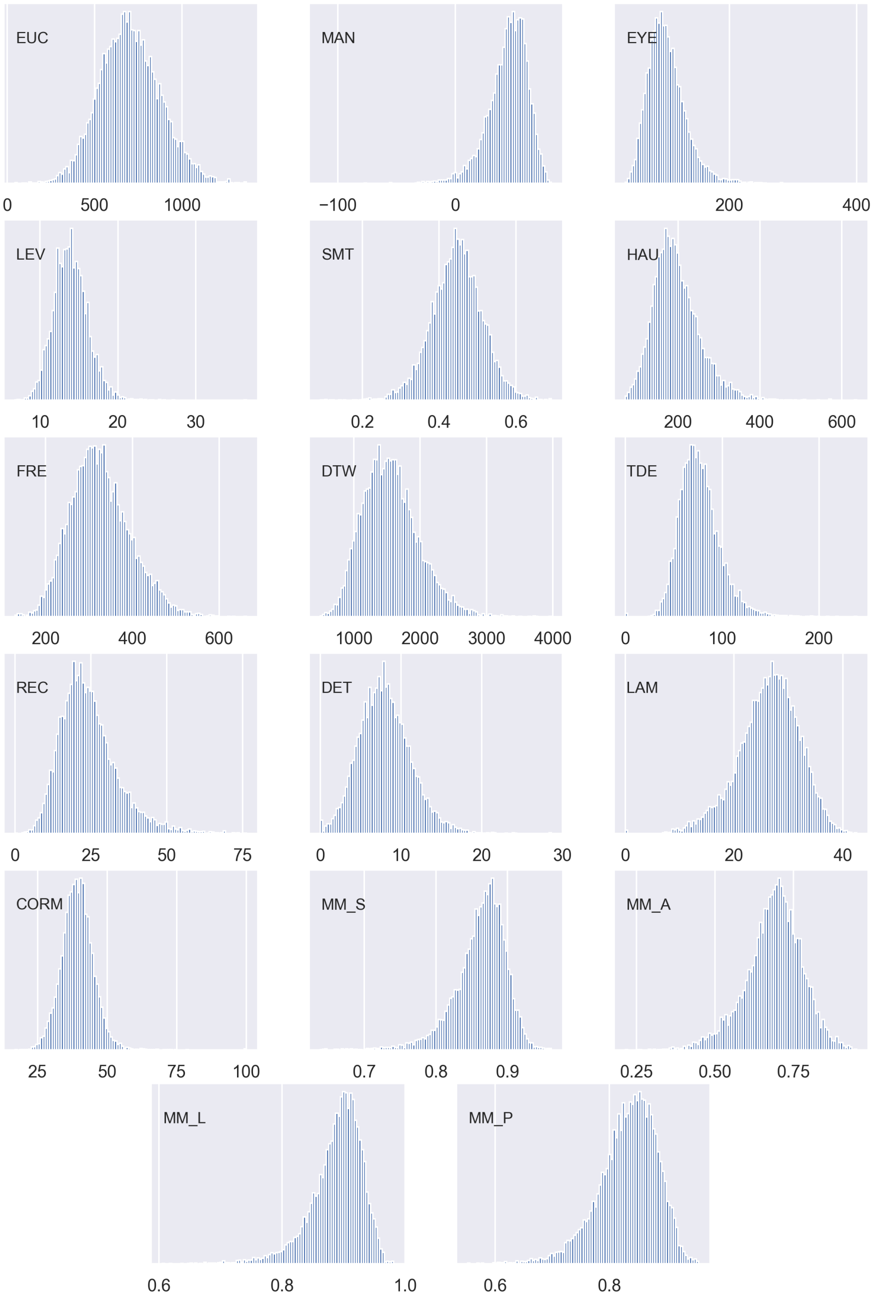


Figure 3.2: Histogram of metrics show a normal distribution.

distributions become very distinct. On the other hand, the standard metrics show a monotonic and linear drop-off revealing the sensitivity to noise and also again hinting at weaker discrimination in the space of comparing scanpaths.

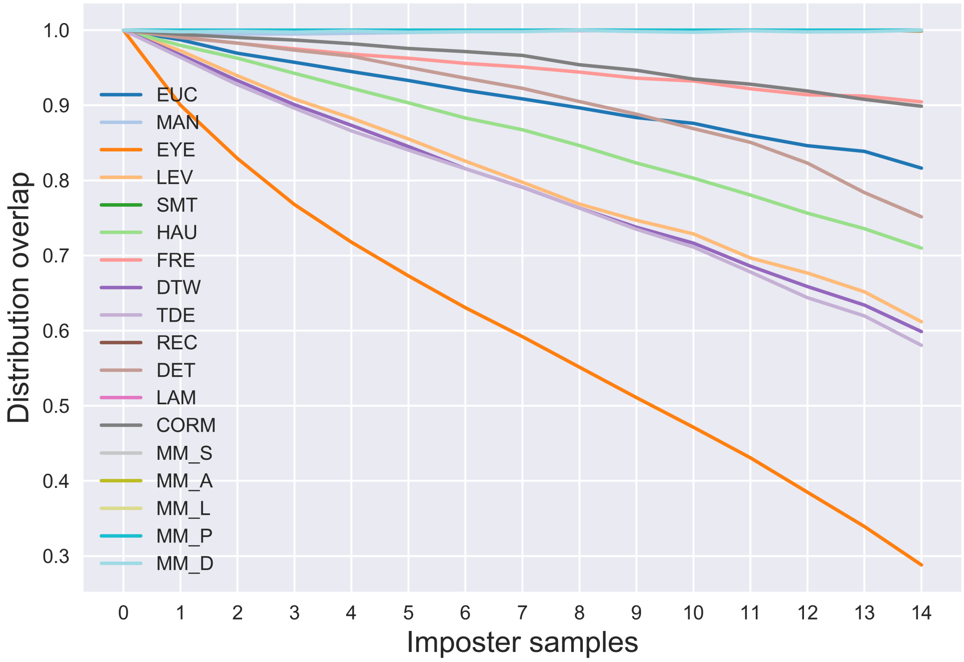


Figure 3.3: Sensitivity to increasing number of imposter samples to the test set in measuring inter-observer distances. This shows the ratio of intersection vs K, the number of imposters introduced to the observer pool.

### 3.2.2 Robustness to Issues in Scanpath Comparison

Taking a principled approach, one may consider various types of perturbations that can occur from one scanpath to another. In a natural setting, one might expect these differences to be predominately driven by differences among viewing patterns across observer or from one image to another. However, the discernibility of differences can be studied in the context of a family of specific perturbations. It is worth noting that some of these perturbations may also occur by virtue of noise in measurement of gaze position, or by virtue of the stochastic nature of gaze points.

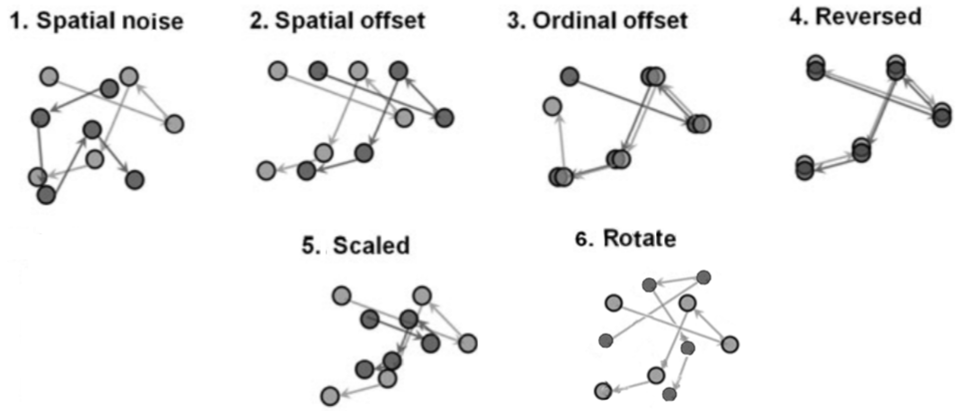


Figure 3.4: A depiction of different conditions that may reveal sensitivity of sequential metrics. These may happen by virtue of small differences in viewing patterns, noise in data capture, or the overall stochastic nature of the process. (Derived from [22])

A set of possible perturbations is shown in figure 3.4. This figure reveals different perturbations of a scanpath that should result in differences in similarity scores, but also provides significant insight into the degree of sensitivity of different measurements to these perturbations. These are as follows:

1. Spatial Noise: The individual fixation locations may be perturbed spatially and independently, while maintaining their overall order.

2. Spatial Offset: Individual fixation locations may be perturbed in a common fashion, where each is subject to the same transformation.

3. Ordinal Offset: The specific ordering may be offset. (e.g. supposing two se-

quences are identical, but one subsequence starts at fixation N and the other at fixation N+1 with equal lengths.

4. Reversed: All fixation locations are identical, but the order they are visited in is reversed.

5. Scaled: The overall geometry of the fixation patterns are identical, but one of the two is subject to global spatial scaling so that the degree of eccentricity of fixations is smaller or larger in the scaled case.

In all of these instances, a metric should be capable of discerning differences, albeit the point of interpretability can be understood much more clearly in the context of observing the degree of sensitivity to such perturbations.

In this regard, we observe the following:

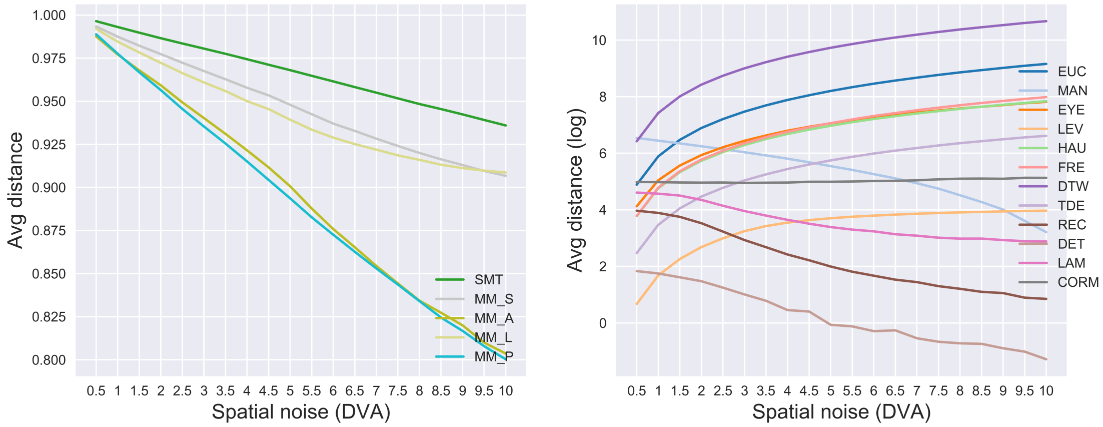


Figure 3.5: Spatial noise: Each fixation has been moved according to a random sample from a Gaussian distribution with $\sigma$ according to degree of visual angle

Figure 3.5 reveals the sensitivity of different metrics to the addition of spatial noise. There is a notable dropoff for increasing spatial offset, that is linear in most

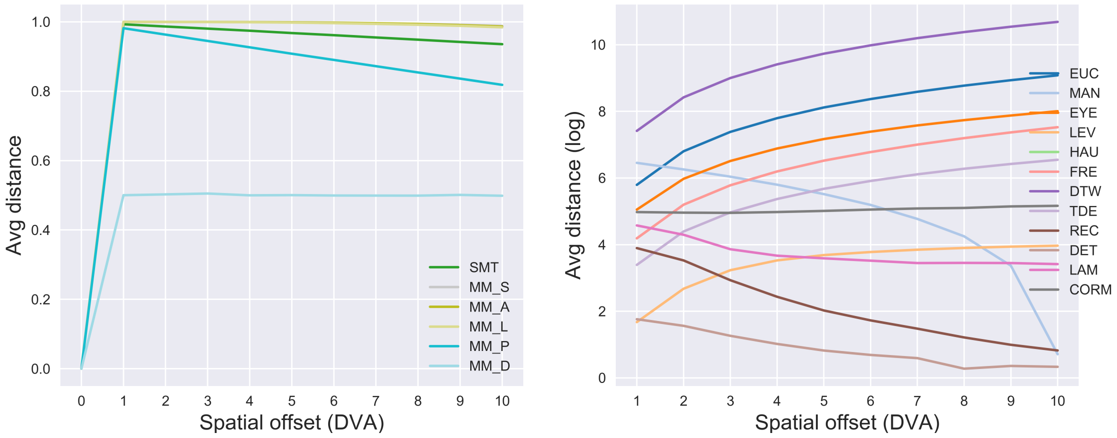


Figure 3.6: Spatial offset: all of fixation have been moved according to a random sample from a Gaussian distribution with $\sigma$ according to degree of visual angle

instances for the ScanMatch and MultiMatch metrics; some metrics show less sensitivity to spatial offset. For other metrics, the behaviour is much more mixed. A family of metrics show common patterns in their behaviour as distance of spatial offset is increased, the average distance produced by the models increases monotonically. Some models fail to adhere to this behaviour. For the case of "tied" spatial offset where all points are subject to a common transformation, the family of metrics that are well behaved for random perturbations show the same well-behaved monotonic increase in inter-scanpath distance.

Figure 3.7 depicts sensitivity to ordinal offset. As the degree of ordinal offset increases, one observes a decrease for the ScanMatch and MultiMatch metrics and virtually no change for SMT, HAU and EYE metrics. In the case of decrease, this follows a inverse sigmoid profile. For other metrics (in particular those that are well-behaved with respect to spatial perturbations), there is consistency in an increase in distance subject to spatial offset, which is followed by a decrease for larger ordinal

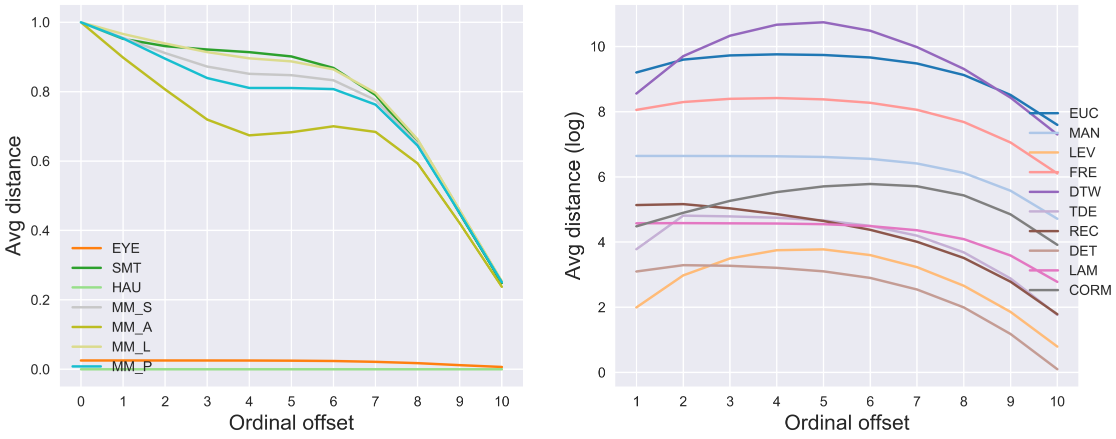


Figure 3.7: Ordinal offset: average distance when scanpaths have temporally shifted in clockwise

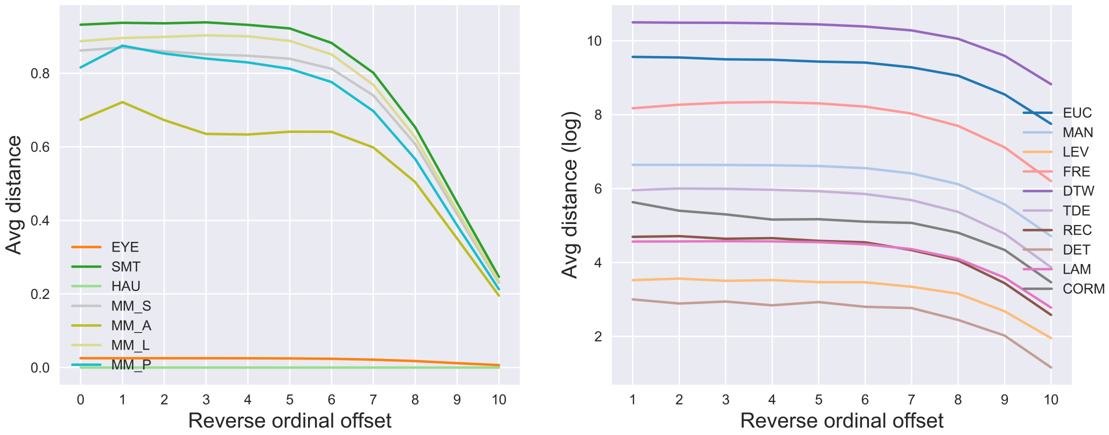


Figure 3.8: Reverse Ordinal offset: average distance when scanpaths have temporally shifted in counter-clockwise

offsets. Note that this is a byproduct of the number of sequences that are sufficiently long to allow such an ordinal offset to be considered, which is much higher for 5 than for any of the cases beyond this. The reversed ordinal offset shown in Figure 3.8 demonstrates sensitivity to distance subject to a reversal of the sequence. In this case, virtually all of the metrics show a relatively small degree of sensitivity to reversal albeit this becomes quite pronounced for the higher reverse ordinal offset. It is interesting to note that the SPP results (that appear in the appendix) don't show the same sensitivity beyond a defined length, because of the minimum rule used in calculating their distance.

Overall, this analysis reveals some interesting characteristics of the sensitivity of distances to different types of perturbations. In some instances, scores are quite erratic, or vary in a manner that is counterintuitive. In other cases, models appear well behaved, and we also seem some degree of correlation appearing at a qualitative level among how these distances vary as the degree of correlation is varied. Models that seem to show the most sensitivity include EYE, DTW, DET and to a lesser extent LEV, HAU, TDE. However, EYE and DET seem to show weakness to ordinal offset, the former by construction.

## 3.3 What is Captured by Different Metrics?

Most metrics proposed have a history that derives from measuring similarity of trajectories, or have roots in sequential comparisons which tend to imply metrics that cast trajectories as strings. While some comparisons have been made among such metrics [3], there is not yet a consensus on what makes for a suitable metric. Moreover, little attention has been paid to the intuitive interpretation of these metrics or their qualitative behaviour.

Figure-3.9 and Figure- 3.10 show for a variety of metrics a reference scanpath in green, and a number of sequences from other observers ranked from most similar to least similar (left to right) each according to a specific metric. The specific metrics shown are referenced in the figure caption. As this figure reveals, there is some

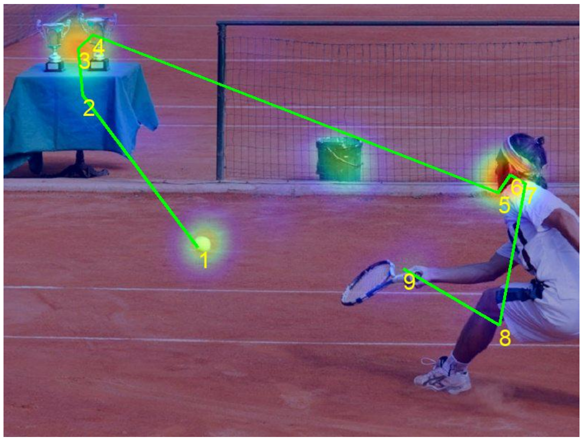
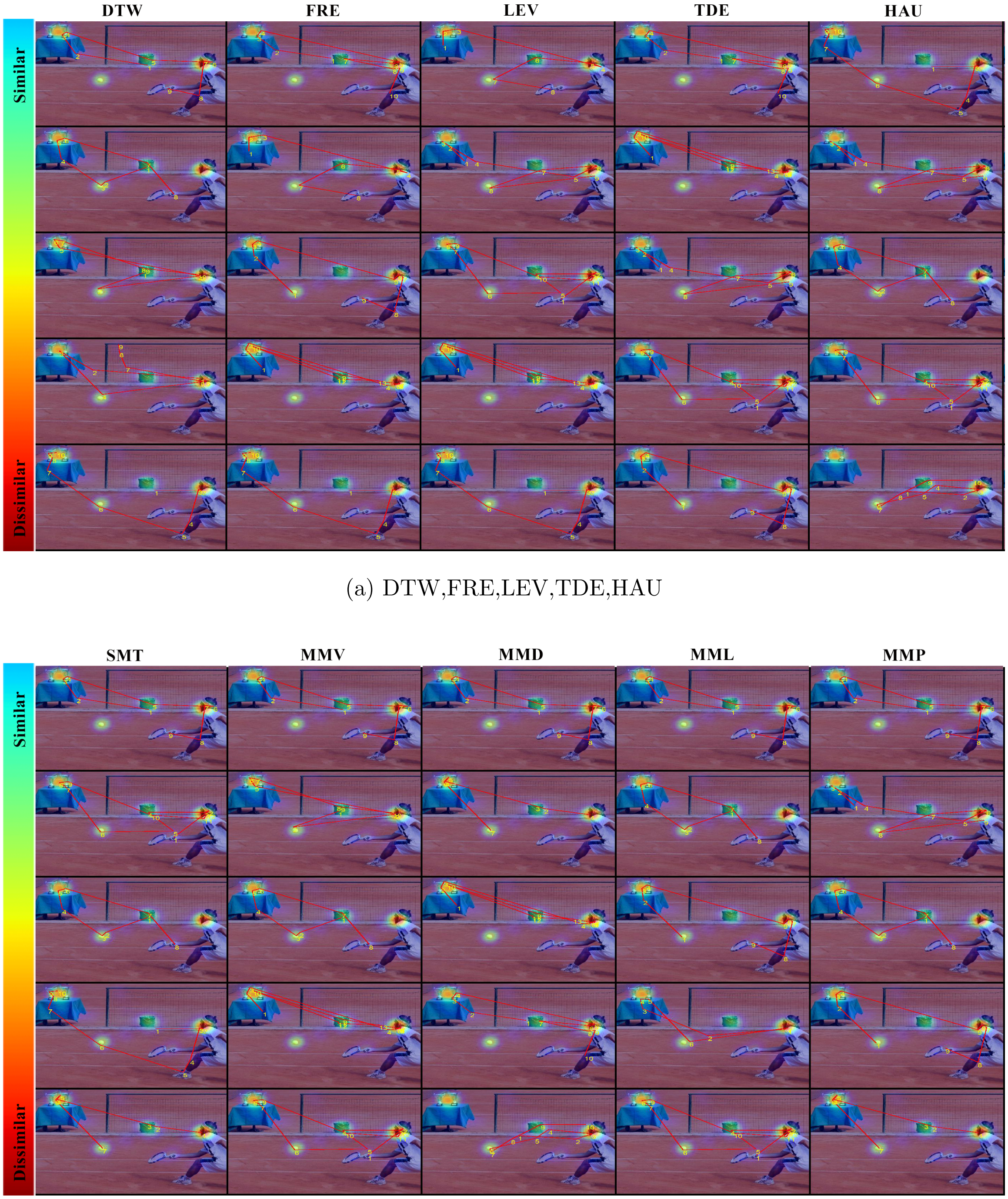


Figure 3.9: Base scanpath

intuitive sense that at least a subset of these metrics is representative of the degree of agreement between pairs of scanpaths. It is not clear however how to discern which metric is most useful. Moreover, it can be seen that scanpaths can be highly divergent even if there is general agreement on a few items in the scene lending support to the argument put forth for metrics that capture *scanpath plausibility* (SPP).

Given that there are a variety of metrics, one might also surmise that some of these may carry similar information, while others may differ significantly in their characterization of distance. One way of considering this directly is by examining the correlation structure among different metrics across a wide array of saliency models and images to determine which produce similar model rankings, or more specifically, to what extent is the structure of relative model rankings similar. To this end, we examine a variety of different saliency algorithms and quantify the Pearson correlation among different metrics across algorithms. This analysis appears in Table A. Note that the balance of this paper provides specific recommendations for metrics and corresponding justification for their use and for this reason, the specific saliency

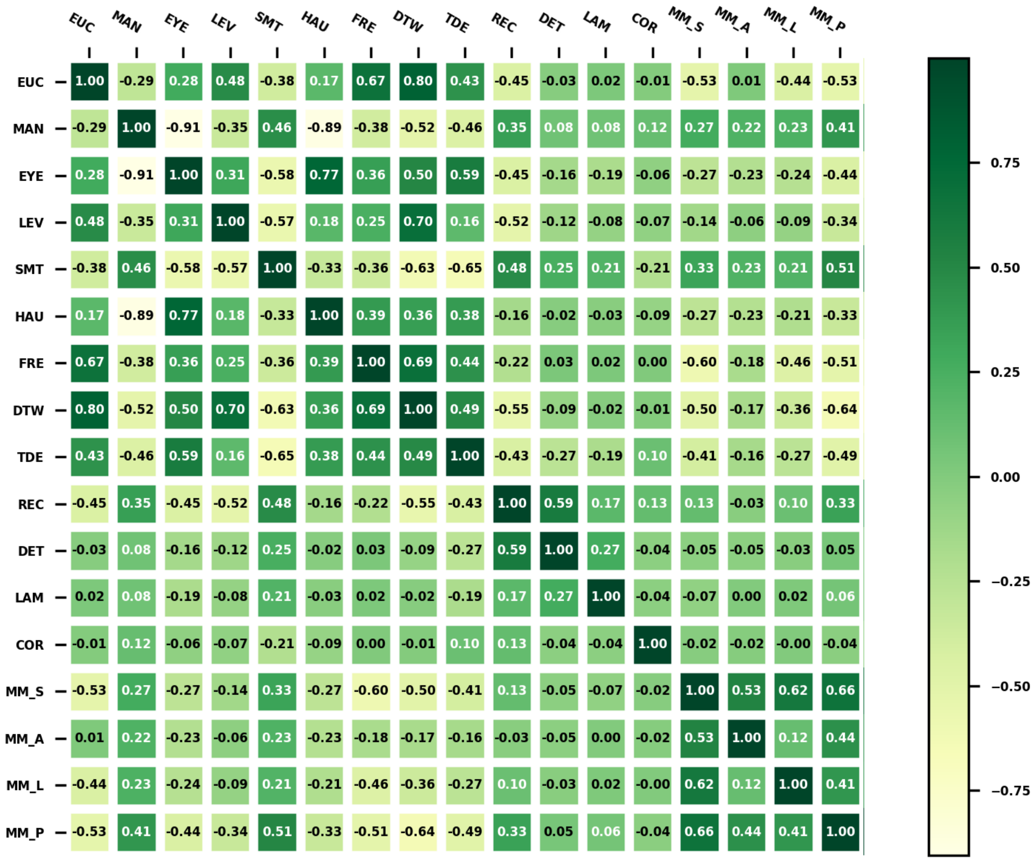


(a) DTW,FRE,LEV,TDE,HAU

(b) SMT, MultiMatch family

Figure 3.10: Comparison between a reference scanpath (green) and scanpaths from other observers (red). Observer scanpaths are shown based on degree of similarity to the reference (left: most similar to right: least similar). Each column corresponds to a different metric that characterizes the degree of similarity (Top to Bottom: DTW, Frechet, Levinshtein, TDE, Hausdorff).

models considered, and an associated benchmark of their performance based on our proposed methods are deferred to the latter part of section 3.5.



| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | COR | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EUC | 1.00 | -0.29 | 0.28 | 0.48 | -0.38 | 0.17 | 0.67 | 0.80 | 0.43 | -0.45 | -0.03 | 0.02 | -0.01 | -0.53 | 0.01 | -0.44 | -0.53 |
| MAN | -0.29 | 1.00 | -0.91 | -0.35 | 0.46 | -0.89 | -0.38 | -0.52 | -0.46 | 0.35 | 0.08 | 0.08 | 0.12 | 0.27 | 0.22 | 0.23 | 0.41 |
| EYE | 0.28 | -0.91 | 1.00 | 0.31 | -0.58 | 0.77 | 0.36 | 0.50 | 0.59 | -0.45 | -0.16 | -0.19 | -0.06 | -0.27 | -0.23 | -0.24 | -0.44 |
| LEV | 0.48 | -0.35 | 0.31 | 1.00 | -0.57 | 0.18 | 0.25 | 0.70 | 0.16 | -0.52 | -0.12 | -0.08 | -0.07 | -0.14 | -0.06 | -0.09 | -0.34 |
| SMT | -0.38 | 0.46 | -0.58 | -0.57 | 1.00 | -0.33 | -0.36 | -0.63 | -0.65 | 0.48 | 0.25 | 0.21 | -0.21 | 0.33 | 0.23 | 0.21 | 0.51 |
| HAU | 0.17 | -0.89 | 0.77 | 0.18 | -0.33 | 1.00 | 0.39 | 0.36 | 0.38 | -0.16 | -0.02 | -0.03 | -0.09 | -0.27 | -0.23 | -0.21 | -0.33 |
| FRE | 0.67 | -0.38 | 0.36 | 0.25 | -0.36 | 0.39 | 1.00 | 0.69 | 0.44 | -0.22 | 0.03 | 0.02 | 0.00 | -0.60 | -0.18 | -0.46 | -0.51 |
| DTW | 0.80 | -0.52 | 0.50 | 0.70 | -0.63 | 0.36 | 0.69 | 1.00 | 0.49 | -0.55 | -0.09 | -0.02 | -0.01 | -0.50 | -0.17 | -0.36 | -0.64 |
| TDE | 0.43 | -0.46 | 0.59 | 0.16 | -0.65 | 0.38 | 0.44 | 0.49 | 1.00 | -0.43 | -0.27 | -0.19 | 0.10 | -0.41 | -0.16 | -0.27 | -0.49 |
| REC | -0.45 | 0.35 | -0.45 | -0.52 | 0.48 | -0.16 | -0.22 | -0.55 | -0.43 | 1.00 | 0.59 | 0.17 | 0.13 | 0.13 | -0.03 | 0.10 | 0.33 |
| DET | -0.03 | 0.08 | -0.16 | -0.12 | 0.25 | -0.02 | 0.03 | -0.09 | -0.27 | 0.59 | 1.00 | 0.27 | -0.04 | -0.05 | -0.05 | -0.03 | 0.05 |
| LAM | 0.02 | 0.08 | -0.19 | -0.08 | 0.21 | -0.03 | 0.02 | -0.02 | -0.19 | 0.17 | 0.27 | 1.00 | -0.04 | -0.07 | 0.00 | 0.02 | 0.06 |
| COR | -0.01 | 0.12 | -0.06 | -0.07 | -0.21 | -0.09 | 0.00 | -0.01 | 0.10 | 0.13 | -0.04 | -0.04 | 1.00 | -0.02 | -0.02 | -0.00 | -0.04 |
| MM_S | -0.53 | 0.27 | -0.27 | -0.14 | 0.33 | -0.27 | -0.60 | -0.50 | -0.41 | 0.13 | -0.05 | -0.07 | -0.02 | 1.00 | 0.53 | 0.62 | 0.66 |
| MM_A | 0.01 | 0.22 | -0.23 | -0.06 | 0.23 | -0.23 | -0.18 | -0.17 | -0.16 | -0.03 | -0.05 | 0.00 | -0.02 | 0.53 | 1.00 | 0.12 | 0.44 |
| MM_L | -0.44 | 0.23 | -0.24 | -0.09 | 0.21 | -0.21 | -0.46 | -0.36 | -0.27 | 0.10 | -0.03 | 0.02 | -0.00 | 0.62 | 0.12 | 1.00 | 0.41 |
| MM_P | -0.53 | 0.41 | -0.44 | -0.34 | 0.51 | -0.33 | -0.51 | -0.64 | -0.49 | 0.33 | 0.05 | 0.06 | -0.04 | 0.66 | 0.44 | 0.41 | 1.00 |

Table 3.1: The Spearman rank correlation among different metrics across algorithms

Note that some metrics (e.g. Levenshtein distance) are poorly correlated with others while others show a greater degree of agreement. With that said, it remains unclear what power these metrics carry in quantifying how well a pair of scanpaths match or how well a predicted scanpath matches a set of ground truth exemplars.

## 3.4 Scanpath Plausibility: Towards a More Intuitive Interpretation

Consider a scene with $k$ regions that tend to be gazed at. There are several possibilities in this scenario; it is possible that some of the regions/objects are gazed at preferentially and tend to be visited in some particular order. In the extreme case, this order may be random. At the other extreme, every observer might assume the same specific order. The reality for most images is likely to be somewhere intermediate to this. E.g. Most observers might first fixate one item before shifting to a second with equal probability and some preference. Alternatively, there may be a rank ordering built into selections that are made coupled with some randomness in this order. All of these considerations present challenges for metrics. Moreover, as an image becomes more complex and contains more regions that tend to be focal points, combinatorics may imply an explosion in the degree of variability among observers (given a fixed number of focal points, there is a combinatorial explosion of sequences as the number of focal points increases) even if the amount of randomness is small.

It is evident from this line of thinking that a metric that uses averaging may be inappropriate. If N observers produce $\approx M$ different strategies for viewing an image, and a model is successful in reproducing one of these strategies, this still implies that $\frac{M-1}{M}$ of instances included in the metric are in some sense noise. The nature of the problem when trying to predict *average* behaviour inherently struggles with the problem of plausible patterns being drowned out by the large number of equally plausible alternatives.

With this in mind, we propose a set of metrics deemed ScanPath Plausibility (SPP). The heart of SPP is to measure the min distance between a model's output sequence, and the set of observer scanpaths. This may be coupled with any metric, but we specifically propose a few appealing options $SPP_{DTW}$, and $SPP_{TDE}$ for reasons discussed in the balance of this section with particular emphasis placed on $SPP_{DTW}$.

## Is SPP More Useful in Determining Model Performance?

In section- 3.2, a decision theoretic definition was put forth for measuring how discriminative two models are with respect to a metric. The crux of this, is that if a metric is applied to two different populations of data (e.g. viewers of the same image vs. different images) each of these implies a distribution of distances for the within-image and different-image comparison. For example, if N-1 observers are measured in their distance to 1 observer N times in a leave one out fashion, this can be done for either the 1 left out, or a random imposter from another image. In this manner, we are able to measure the extent to which a metric can discern whether two viewers examined the same image or different images. Results from this experiment are shown in Table 3.2. In particular, the comparison of the pool of observers with the left-out sample elicits a distribution, for both within-class and imposter cases. The area of overlap between these two distributions is a useful measure of the degree of confusion one might expect, where 0 implies no overlap among distances. In this instance, it is clear that there is a significant advantage to some metrics, but also to the SPP variant in particular.

In addition to revealing behaviour of metrics the results of this experiment for SPP

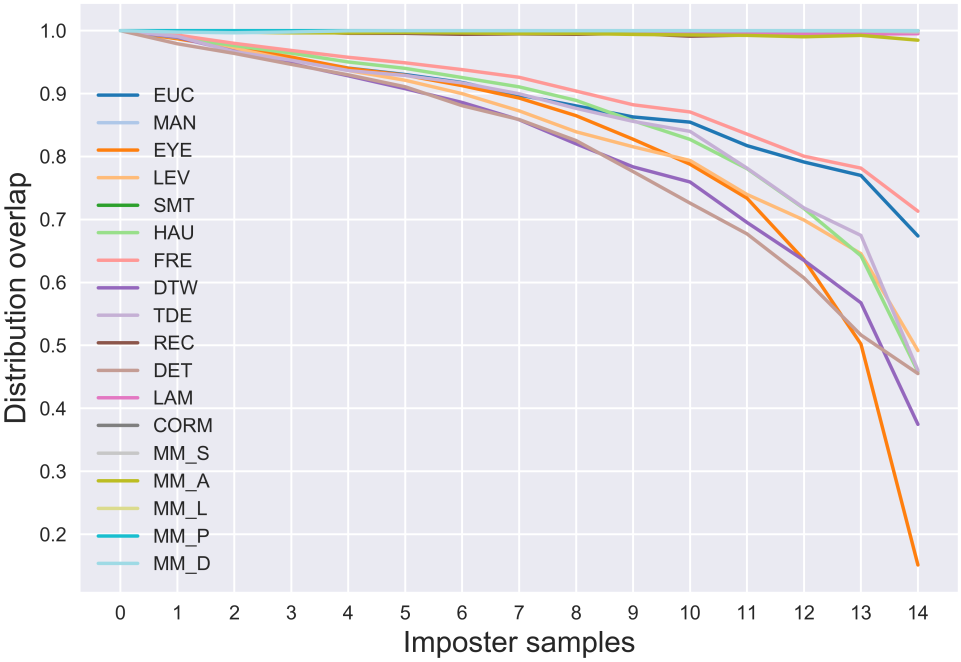


Figure 3.11: Sensitivity of pollution of the observer pool by an increasing number of imposter samples. This effectively measures inter-observer similarity for a pool of $N-k$ observers where $k$ are sequences chosen at random from observers of a different image. $k$ is gradually increased to examine the degree of sensitivity to noise samples.

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MEAN | 0.82 | 1.00 | 0.29 | 0.61 | 1.00 | 0.71 | 0.90 | 0.60 | 0.58 | 1.00 | 0.75 | 1.00 | 0.90 | 1.00 | 1.00 | 1.00 | 1.00 |
| SPP | 0.67 | 1.00 | 0.15 | 0.49 | 1.00 | 0.46 | 0.71 | 0.37 | 0.46 | 1.00 | 0.46 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 |

Table 3.2: Area of intersection for two distributions. One is based on distances of observers viewing the same image, and the other is between images. In practice, a strong metric should elicit a very different distribution.

are unequivocal. The SPP distances show significantly higher capacity for separating these classes in part due to resilience to domain relevant noise that is intrinsic to inter-observer variability, but also given the high intrinsic variability of scanpaths that implies a large distance when an imposter is considered.

To take this analysis further, we ask a different question. Suppose inter-observer scanpath similarity is considered for a given image, and this is compared with inter-observer scanpath similarity where an increasing number of imposters is introduced to the observer pool. In this case, leave one out analysis is performed on the group for all N-1 observers (with $K < N - 1$ imposters). One can also consider how the distributions diverge as the observer pool is increasingly polluted with random scanpaths from other images. The results of this analysis appear in Figure 3.11. One can see that the SPP (Table A.4) based on a min distance across observers is much less sensitive to noisy samples being introduced. Moreover when the sample is entirely or almost entirely imposters, the distributions become very distinct. Alternatively, the standard metrics show a monotonic and somewhat linear drop-off revealing the sensitivity to noise and also again hinting at weaker discrimination in the space of comparing scanpaths.

This confirms the observation that more separation can be observed based on $SPP$, and that traditional metrics would lead one to conclude that the 2 sets of observations are identical for many metrics. At first glance, one might be inclined to rely on the EYE metric for it's separability and behaviour as imposters are introduced. However, it is notable that ordinal perturbations are irrelevant. For this reason, one might consider using both $DTW_{SPP}$ and $EYE$ in concert to measure model performance or compare populations of observers.

## 3.5 A Sequential Model Benchmark

In this section, we leverage the observations made to date in order to assess existing models of saliency or gaze prediction through the lens of sequential analysis. The number of models that are sequential by design is extraordinarily small compared with those that consider a spatial focus of attention. For this reason, we have applied a simple method for simulating scanpaths from static saliency maps following classic work in saliency [37].

In this analysis, static methods that elicit a heatmap-style representation of saliency are sampled based on their strongest point (this is known as Winner-Take-All in prior work). Subsequently, the sampled region is suppressed and the next highest point is selected in order to simulate a scanpath from the static sample set. Given that the models are relatively capable of identifying the strongest regions of interest to a human observer, this approach tends to produce as output a sequence in some order between the top several locations. In this fashion, we can consider what base spatial saliency models have a static mapping that lends itself well to sequences consistent with human observers, or consider the limited set of models that are inherently sequential by design (LeMeur,PathGAN).

Table 3.3 reveals benchmark scores for an array of models induced to produce scanpaths, or those that produce scanpaths natively in comparison to human scanpaths for the same images. Importantly, this presents the first comprehensive demonstration of a means of comparing scanpaths across different models, experimental conditions, images or populations that is sequential. Moreover, it establishes what metrics make most sense, their sensitivity, and capacity to successfully separate observations that derive from different conditions. Somewhat surprising is the effectiveness of the WTA mechanism given a base representation of saliency. For the more discernible metrics, it is clear that the LeMeur approach fares slightly better, however the recent PathGAN is significantly worse on most accounts.

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AIM | 831.78 | 23.27 | 168.98 | 15.35 | 0.76 | 242.37 | 386.23 | 2032.22 | 113.64 | 4.33 | 0.66 | 8.27 | 40.86 | 0.84 | 0.67 | 0.86 | 0.77 |
| AWS | 828.08 | 23.48 | 164.34 | 15.11 | 0.76 | 244.82 | 383.03 | 2010.82 | 113.07 | 4.90 | 0.83 | 9.25 | 40.15 | 0.84 | 0.67 | 0.86 | 0.77 |
| CAS | 761.36 | 24.29 | 167.34 | 15.19 | 0.77 | 230.77 | 353.70 | 1910.34 | 102.37 | 4.63 | 0.93 | 7.94 | 41.36 | 0.85 | 0.66 | 0.88 | 0.79 |
| CVS | 898.08 | 1.83 | 235.55 | 16.07 | 0.76 | 291.37 | 413.44 | 2353.33 | 128.56 | 2.70 | 0.45 | 6.42 | 42.24 | 0.84 | 0.67 | 0.86 | 0.76 |
| DVA | 847.99 | 19.93 | 176.01 | 15.48 | 0.76 | 251.78 | 392.54 | 2088.77 | 116.76 | 4.13 | 0.70 | 8.35 | 39.27 | 0.84 | 0.67 | 0.86 | 0.76 |
| GBVS | 707.28 | 32.96 | 147.79 | 14.85 | 0.77 | 202.24 | 323.15 | 1730.56 | 94.50 | 5.64 | 1.13 | 8.41 | 41.84 | 0.86 | 0.67 | 0.89 | 0.81 |
| IKN | 777.06 | 25.81 | 164.95 | 15.20 | 0.77 | 228.45 | 358.76 | 1913.76 | 105.72 | 4.47 | 0.75 | 7.73 | 41.67 | 0.85 | 0.67 | 0.88 | 0.79 |
| IMSIG | 796.70 | 23.28 | 166.60 | 15.12 | 0.77 | 236.84 | 370.47 | 1959.85 | 105.25 | 5.01 | 0.94 | 8.72 | 40.49 | 0.85 | 0.67 | 0.87 | 0.78 |
| QSS | 909.77 | 9.17 | 200.28 | 15.77 | 0.76 | 287.47 | 424.04 | 2267.87 | 130.19 | 3.77 | 0.65 | 8.09 | 40.51 | 0.84 | 0.67 | 0.86 | 0.76 |
| SSR | 874.44 | 15.22 | 188.71 | 15.54 | 0.76 | 262.36 | 408.29 | 2159.36 | 120.03 | 4.16 | 0.76 | 7.98 | 40.24 | 0.84 | 0.67 | 0.87 | 0.77 |
| SUN | 850.72 | 16.39 | 185.09 | 15.59 | 0.76 | 256.88 | 393.81 | 2117.18 | 117.92 | 3.83 | 0.69 | 8.13 | 40.05 | 0.84 | 0.67 | 0.86 | 0.76 |
| cG | 599.08 | 18.56 | 187.09 | 15.11 | 0.78 | 220.50 | 286.53 | 1693.72 | 87.20 | 4.46 | 1.14 | 7.03 | 45.62 | 0.85 | 0.49 | 0.85 | 0.81 |
| SAM-VGG | 728.45 | 47.33 | 106.02 | 14.03 | 0.77 | 187.48 | 340.53 | 1600.90 | 94.71 | 6.10 | 1.06 | 10.60 | 39.63 | 0.86 | 0.70 | 0.88 | 0.82 |
| OpenSalicon | 730.94 | 41.46 | 120.34 | 14.19 | 0.77 | 200.00 | 348.58 | 1639.66 | 91.99 | 6.57 | 1.11 | 10.83 | 38.31 | 0.85 | 0.69 | 0.88 | 0.82 |
| SALGAN | 1112.63 | -14.12 | 213.76 | 15.66 | 0.76 | 424.14 | 555.35 | 2611.46 | 171.14 | 4.92 | 0.81 | 11.04 | 33.30 | 0.83 | 0.69 | 0.83 | 0.77 |
| PathGAN | 1218.31 | -80.62 | 443.05 | 17.52 | 0.73 | 414.35 | 552.11 | 3711.61 | 173.52 | 0.17 | 0.17 | 8.10 | 32.01 | 0.67 | 0.48 | 0.61 | 0.60 |
| LeMeur | 427.38 | 109.58 | 116.62 | 12.48 | 0.57 | 128.85 | 182.75 | 1246.36 | 56.40 | 0.01 | 0.00 | 0.00 | 12.18 | 0.66 | 0.12 | 0.63 | 0.50 |

Table 3.3: A benchmark of saliency models that considers trajectories rather than spatial distributions subject to a variety of different metrics.

## 3.6 Summary

In this chapter we have examined the problem of scanpath prediction with an emphasis on suitable metrics for comparing scanpaths. In doing so, we have demonstrated the surprising capability of simple winner-take-all mechanisms built atop static saliency maps in eliciting reasonable approximations of human scanpaths. Moreover, we have shown how metrics that have traditionally been used in this domain lack an intuitive interpretation, and also provide weak contrast for revealing differences in model capabilities. We have provided careful analysis of different metrics and their characteristics and proposed a viable alternative that provides a stronger measure for model-human or model-model distances. Finally, we provide some specific recommendations for metrics and considerations that are important in moving forward.

# Chapter 4

# Secrets of Sequential Analysis

Comparisons using spatial metrics are a core part of the literature involving gaze prediction. However, as alluded to in earlier chapters, there is reason to be optimistic that sequential analysis might show more insight regarding the data, including revealing differences that are not observable through traditional spatial analysis.

With this consideration in mind, this chapter is dedicated to application of the body of work we earlier developed surrounding metrics to a variety of studies of gaze patterns. As the results and discussion demonstrate, the richer representation afforded by sequential data produces a different interpretation of data than has been revealed prior using traditional spatial analysis.

## 4.1 Stimuli Effect

Among studies that consider gaze data, there are a few including one larger dataset (CAT2000) that is grouped according to stimulus category. An interesting question that arises in the context of sequential analysis is as follows: *Do sequential metrics provide a good degree of resolution in revealing general differences in the similarity of viewing patterns for different categories?*

This question has been addressed directly in applying the set of metrics we have considered to the variety of stimulus categories in CAT2000. The results of this analysis are presented in Table 4.1.

It is interesting to see that there is a wide spread in inter-observer distance as a function of category. More abstract scene representations including fractals, line drawings and inverted scenes seem to elicit less similar viewing patterns. More standard imagery shows an intermediate range of similarity, while simple images including simple patterns or noise show a high degree of similarity. In this latter case, this is likely owing to a relative lack of structure in viewing or eccentricity for noise, and simple viewing strategy for basic patterns. These observations are made quite palpable by sequential gaze analysis whereas more simplistic analysis based on spatial distributions don't reveal such significant contrast across viewing conditions.

## 4.2 Resolution

Another experimental parameter that may be varied systematically is the resolution at which the image is presented. For example, a common size for presenting the image may be used while in some instances this is an upsampled version of a coarse grained image. The effect of varying stimuli resolution and viewing behaviour has been assessed in [41] based on spatial analysis. The authors have performed an eye tracking study by showing images at different resolutions to a set of non-overlapping

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Affective | 1554.69 | 32.17 | 200.55 | 30.73 | 0.40 | 429.68 | 571.34 | 5638.05 | 113.33 | 4.00 | 2.75 | 13.08 | 34.18 | 0.90 | 0.70 | 0.91 | 0.86 |
| BlackWhite | 1468.51 | 31.68 | 210.74 | 30.06 | 0.39 | 431.53 | 558.50 | 5470.13 | 116.43 | 3.59 | 2.34 | 11.85 | 34.55 | 0.90 | 0.70 | 0.91 | 0.86 |
| Fractal | 1632.23 | 26.11 | 227.07 | 31.16 | 0.37 | 452.68 | 598.80 | 6069.34 | 127.71 | 3.16 | 2.10 | 11.34 | 34.82 | 0.89 | 0.70 | 0.90 | 0.84 |
| Inverted | 1703.50 | 25.84 | 225.79 | 31.94 | 0.37 | 442.19 | 595.13 | 6287.62 | 129.76 | 2.58 | 1.91 | 10.63 | 34.67 | 0.89 | 0.71 | 0.91 | 0.84 |
| Line Drawing | 1683.26 | 25.44 | 229.55 | 32.54 | 0.35 | 441.17 | 593.74 | 6444.01 | 133.39 | 2.60 | 1.75 | 10.56 | 34.89 | 0.89 | 0.69 | 0.91 | 0.84 |
| Noisy | 1208.06 | 33.65 | 220.39 | 27.44 | 0.36 | 440.41 | 536.13 | 4771.37 | 116.06 | 4.22 | 2.06 | 11.79 | 36.71 | 0.89 | 0.65 | 0.90 | 0.86 |
| Outdoor | 1513.83 | 30.19 | 212.57 | 30.72 | 0.38 | 438.72 | 581.14 | 5697.78 | 118.53 | 3.50 | 2.21 | 11.38 | 34.89 | 0.89 | 0.70 | 0.90 | 0.85 |
| Pattern | 1375.97 | 29.68 | 223.07 | 29.57 | 0.37 | 445.03 | 558.78 | 5400.19 | 120.59 | 3.66 | 2.03 | 11.42 | 35.92 | 0.89 | 0.67 | 0.91 | 0.85 |
| Satellite | 1547.59 | 26.45 | 231.41 | 30.17 | 0.36 | 454.31 | 588.80 | 5813.68 | 128.96 | 3.06 | 1.96 | 11.11 | 34.96 | 0.89 | 0.69 | 0.91 | 0.85 |
| Social | 1463.49 | 37.10 | 187.82 | 30.25 | 0.41 | 396.76 | 536.32 | 5327.27 | 107.14 | 4.09 | 2.73 | 13.00 | 34.43 | 0.90 | 0.71 | 0.92 | 0.87 |
| MEAN | 1515.29 | 29.84 | 216.85 | 30.46 | 0.38 | 437.21 | 571.86 | 5691.83 | 121.17 | 3.44 | 2.19 | 11.63 | 34.96 | 0.89 | 0.69 | 0.91 | 0.85 |
| SPP | 644.87 | 23.21 | 129.99 | 21.38 | 0.16 | 233.79 | 373.51 | 3418.67 | 53.06 | 0.04 | 0.00 | 0.43 | 12.52 | 0.83 | 0.42 | 0.81 | 0.75 |

Table 4.1: Inter observer results per category for CAT2000

observers. The resolutions range from 4x4-512x512 px and images were interpolated to a fixed size before being displayed to observers. The main point of the Judd et al. study is on the consistency of fixations across resolutions. Using spatial methods that are not able to capture sequential dynamics, the consistency remains constant for 32px and above. This experiment may be re-examined in the context of sequential metrics in order to confirm or revisit these findings. To this end, we consider DTW to compare scanpaths of images at 8 different resolutions to one another. This results in an 8x8 set of comparisons and the results were then normalized by the inter-observer distance of the base resolution. Figure 4.2 shows the results of this analysis and it is evident that changes in behaviour continue for 64x64 images and possibly even for the 128x128 case. After this point viewing behaviours starts to become consistent. Nevertheless, this presents another compelling example of the strength of sequential metrics in revealing more above viewing patterns, especially when an appropriate metric is selected.

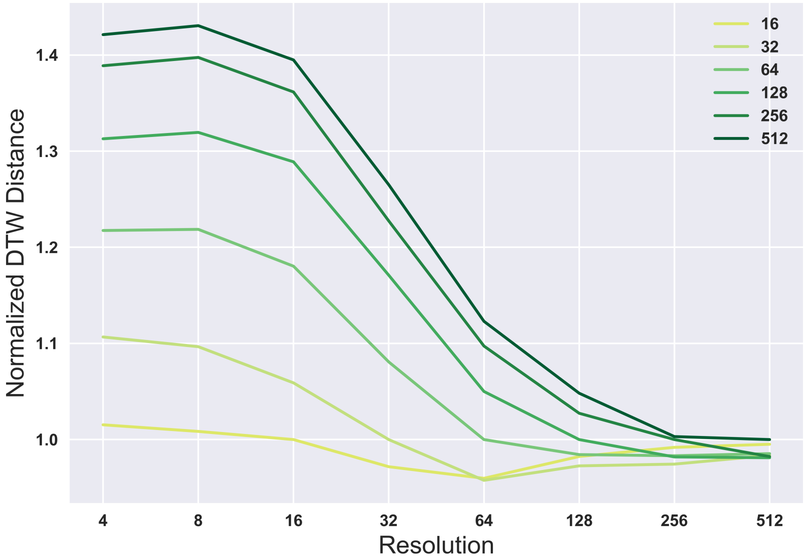


Table 4.2: Inter observer score for different resolutions of an image. Results have been normalized normalized by the inter-observer distance of the base resolution.

## 4.3 Eye-tracking vs Mouse-tracking

With the considerable time and cost of collecting eye tracking data, combined with the need for large-scale data to make use of deep learning models, it has recently become common that saliency models are pre-trained on datasets collected using different proxies for gaze. There a few options to achieve this including mouse-tracking and web-cam based tracking. Salicon [39] is the most used dataset and has been collected through Amazon Mechanical Turk based on mouse-tracking data. In their experiments, to show the effectiveness of their method, they also collected mouse-tracking data for the OSIE dataset for which eye tracking data is available. Given the large number of samples captured by mouse-movements, k-Means clustering is applied to raw mouse coordinates to produce a simulated scanpath using k-means

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| mouse-mouse | 703.29 | 27.26 | 136.40 | 10.75 | 0.77 | 179.95 | 314.08 | 1446.57 | 89.85 | 4.56 | 1.27 | 0.87 | 34.60 | 0.71 | 0.62 | 0.73 | 0.66 |
| mouse-mouse spp | 432.03 | 20.81 | 83.52 | 7.89 | 0.72 | 94.09 | 188.37 | 906.49 | 58.87 | 0.06 | 0.00 | 0.00 | 12.35 | 0.62 | 0.41 | 0.61 | 0.54 |
| Mouse-eye | 824.88 | 8.72 | 213.97 | 15.97 | 0.76 | 256.13 | 370.25 | 2145.35 | 121.03 | 6.45 | 23.08 | 17.64 | 44.92 | 0.83 | 0.66 | 0.85 | 0.76 |
| Mouse-eye SPP | 591.79 | 49.19 | 143.45 | 13.26 | 0.74 | 169.53 | 247.31 | 1528.37 | 85.19 | 4.10 | 22.67 | 13.37 | 18.58 | 0.74 | 0.42 | 0.70 | 0.64 |

Table 4.3: Mouse tracking vs Eye tracking data on OSIE dataset. This shows a remarkable difference in considering within and between class distances, and provides strong evidence for the view that mouse-tracking proxies can't help saccadic models as effectively as static models.

centriods as putative fixation points. The average length of eye tracking scanpaths are 8 fixations for OSIE dataset and therefore we consider 8 centers for simulated scanpaths based on k-means. The results show quite remarkable difference for inter-observer distances between eye-tracking and mouse-tracking data. Table 4.3 shows the results for the mean version of the metrics and also the SPP. This is another case where spatial metrics may show a subtle difference but sequential metrics clearly show a very large substantive difference, suggesting that mouse data is a poor proxy for simulating gaze data.

## 4.4 Saliency in Time

It has been observed that viewing patterns tend to diverge over time with some of the most prominent targets for gaze having been exhausted, or attention disengaged. This analysis is outside of the realm of possibility in considering traditional spatial analysis since there is no notion of time and all fixations are treated equally. In spatial terms, it has been observed that the common pattern of centre-bias expands to a higher degree of eccentricity, but there is relatively little precision exercised in examining how gaze patterns may diverge over longer time periods, and whether metrics have the sensitivity to reveal this.

To this end, we consider per-image inter-observer viewing patterns for the OSIE, restricting sequence lengths as a function of total viewing time. The expectation is that one might see an increase in distance among gaze patterns. Figure 4.1 reveals the expected trend wherein distances increase subject to time up to a critical time-point. Note that each metric is normalized according to its maximum score over the time intervals considered. Moreover, the more sensitive metrics we have identified previously also reveal this contrast much more strongly. Interestingly, some metrics have the characteristic that there is a marginal drop in inter-observer distances after some time, most likely owing to concentric viewing or exhausting of the set of targets of interest. It is interesting to note that in this case DTW shows the expected trend while the EYE model that showed good discriminability runs in contrast to this trend, likely owing to its lack of consideration to order gaze points and associated benefit of having more overall fixations.

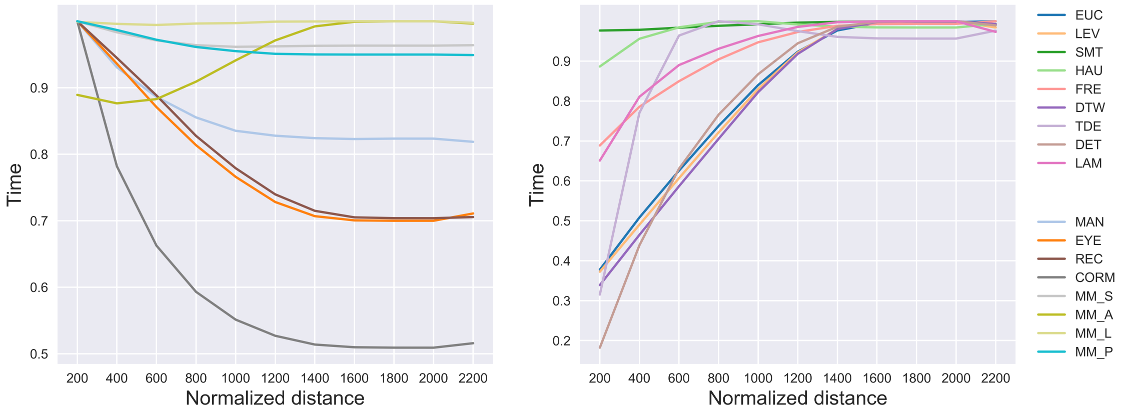


Figure 4.1: Inter observer distance by quantizing according to duration of viewing and normalized by maximum IOC distance per metric

# Chapter 5

# Summary, Limitations and Future Work

In general, simulating and modeling the human visual system has the advantage that it results in human-like behavior. This is beneficial for systems that should interact with humans in a natural manner. Seeking simple solutions to produce an appropriate sequence of spatial selections in the form of fixation points is of value for many human-centric problems as well artificial vision systems that may benefit from active and attentive exploration of a scene.

In this thesis, we have presented a deep analysis of metrics for analyzing human gaze data. In particular, we have focused on the sequential nature of gaze patterns and presented new metrics for analysis of gaze as a sequential process. In doing so, we have also done a critical appraisal of existing metrics in the literature and provided a demonstration of strengths and weaknesses of different approaches. The balance of this chapter summarizes the main contributions / findings of this thesis, and limitations or directions for future work.

## 5.1 Summary

While this thesis focuses on a narrow topic, treatment of this topic requires addressing a substantial breadth of literature, methods, models and carefully considering prevailing thinking in this realm over several decades. This has resulted in a variety of interesting a specific outcomes, which we state in what follows to provide a sense of the big picture that this thesis reveals:

- In Chapter 2, we present a critical appraisal of a very comprehensive set of literature. This is not only at the level of analyzing what has been considered for sequential metrics, but also includes analysis of different proposals that sets the stage for the more technical contributions of Chapter 3.

- There are a vast number of datasets in the literature, but given the heavy focus on spatial analysis of gaze patterns, only few of these are suitable for sequential analysis and discerning this requires careful examination of available data. In the context of this thesis a comprehensive review of available datasets is presented, and those that lend themselves to sequential or spatiotemporal analysis have been identified. Moreover, these have been transformed into a format that allows for training of models, use of our metrics or further work in this domain.

- There is a vast space of metrics, and Chapter 3 presents a series of carefully designed studies to understand these metrics. This begins with identifying what properties one would hope to see from a good metric. Subsequent experiments reveal strengths and weaknesses of different metrics and their sensitivity to different perturbations of the data.

- The results of Chapter 3 identify a few metrics that are especially discriminative in separating gaze patterns that derive from different populations. This notably includes the metric we propose and a few others. Among the others, the thorough nature of experiments reveals some of the caution that may be required in exercising these such as the degree of sensitivity to the ordering of fixation points.

- There are no convincing benchmarks for saliency models that consider the problem from a sequential perspective, and this work presents such a benchmark in addition to a justification for the metrics supporting model superiority.

- The thesis begins with the claim that there should evidently be more information in a sequence than a static representation. In chapter 4, a variety of examples are provided that show that conclusions that are based on spatial analysis alone fail to hold up, or change when viewed through the lens of sequential analysis. Each of these results is important in it's own right, but also as a justification for the careful study of sequential models of gaze in moving forward.

Finally, on the balance of the work presented in this thesis the author hopes that this will inspire others to more deeply consider the intriguing and rich problem of sequential analysis of gaze.

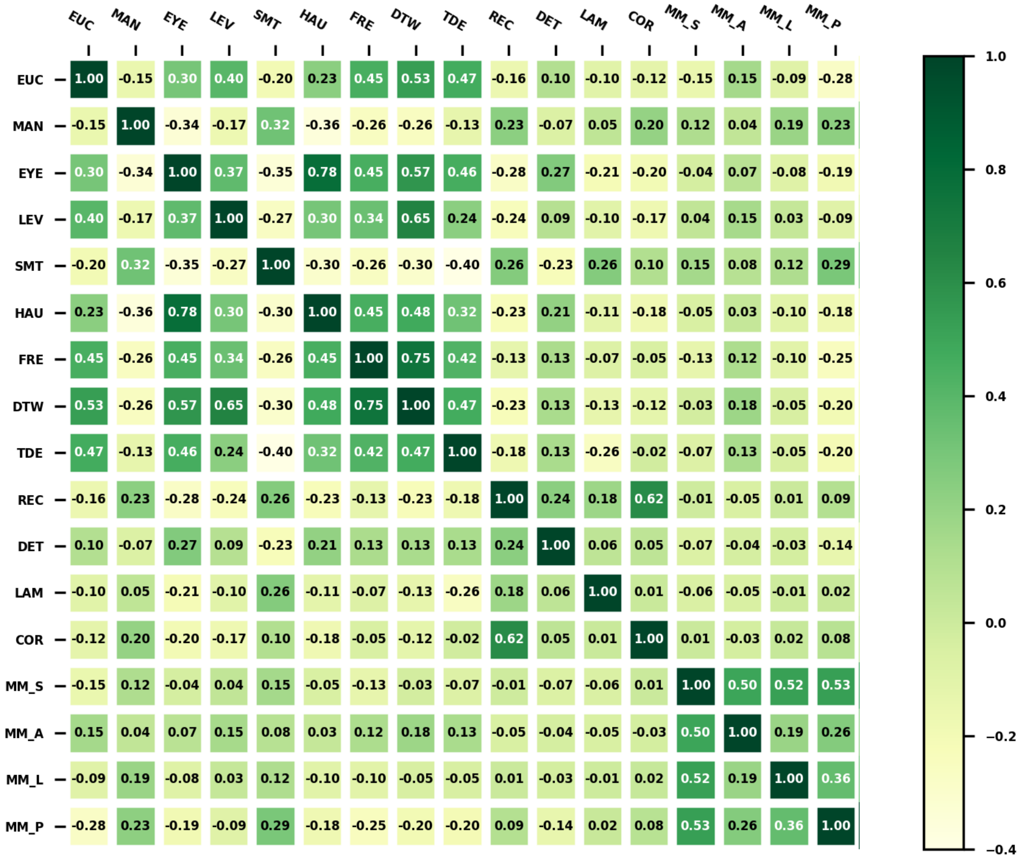
## 5.2 Limitations and Future works

The nature of the domain covered by this thesis implies that analysis could go on forever. There are many problems in the realm of top-down vs. bottom up viewing, the role of emotion, complexity of images and other factors that could be analyzed in much greater detail. There is also a wealth of existing work that authors might choose to revisit in making use of the results of this thesis to provide new perspectives on their own data. Finally, there is a very short history of computational models that address the problem of saliency, attention and gaze sequentially and further exploration of such models is warranted. The work in this thesis provides a foundation for computational modeling efforts in this domain to be explored with confidence in the metrics that support the outcomes that are yet to be revealed.

# Appendix A

# AP1.

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P | MM_D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 703.85 | 44.50 | 101.73 | 13.97 | 0.45 | 196.94 | 326.33 | 1559.30 | 76.91 | 24.34 | 8.02 | 26.48 | 39.39 | 0.86 | 0.69 | 0.89 | 0.83 | 0.50 |
| 1 | 711.85 | 41.39 | 110.30 | 14.18 | 0.43 | 201.49 | 328.95 | 1609.35 | 79.37 | 23.08 | 7.71 | 25.97 | 39.35 | 0.86 | 0.68 | 0.89 | 0.83 | 0.50 |
| 2 | 722.77 | 38.41 | 118.70 | 14.47 | 0.42 | 205.32 | 331.15 | 1666.83 | 82.14 | 21.79 | 7.43 | 25.45 | 39.31 | 0.86 | 0.68 | 0.89 | 0.82 | 0.50 |
| 3 | 730.13 | 35.13 | 127.61 | 14.74 | 0.41 | 209.79 | 333.05 | 1722.08 | 84.53 | 20.64 | 7.15 | 24.83 | 39.32 | 0.86 | 0.67 | 0.88 | 0.82 | 0.50 |
| 4 | 737.48 | 32.13 | 136.09 | 14.96 | 0.40 | 214.14 | 334.86 | 1771.37 | 86.98 | 19.32 | 6.81 | 24.29 | 39.27 | 0.86 | 0.67 | 0.88 | 0.82 | 0.50 |
| 5 | 744.68 | 29.01 | 144.46 | 15.21 | 0.38 | 218.35 | 336.20 | 1822.85 | 89.06 | 18.20 | 6.57 | 23.84 | 39.30 | 0.85 | 0.67 | 0.88 | 0.81 | 0.50 |
| 6 | 752.79 | 25.80 | 152.97 | 15.47 | 0.37 | 222.74 | 337.91 | 1877.78 | 91.14 | 17.17 | 6.35 | 23.32 | 39.31 | 0.85 | 0.66 | 0.88 | 0.81 | 0.50 |
| 7 | 759.13 | 22.99 | 161.08 | 15.72 | 0.36 | 225.95 | 338.95 | 1923.53 | 93.17 | 16.17 | 6.10 | 22.65 | 39.45 | 0.85 | 0.66 | 0.88 | 0.80 | 0.50 |
| 8 | 765.75 | 19.82 | 169.88 | 15.98 | 0.35 | 230.27 | 340.36 | 1975.00 | 95.57 | 15.17 | 5.80 | 21.91 | 39.94 | 0.85 | 0.65 | 0.88 | 0.80 | 0.50 |
| 9 | 772.88 | 16.74 | 178.57 | 16.16 | 0.33 | 234.68 | 342.07 | 2021.15 | 97.93 | 13.99 | 5.44 | 21.31 | 40.07 | 0.85 | 0.65 | 0.88 | 0.79 | 0.50 |
| 10 | 776.90 | 13.88 | 186.98 | 16.32 | 0.32 | 238.56 | 342.81 | 2059.89 | 99.91 | 12.83 | 5.02 | 20.55 | 40.41 | 0.85 | 0.65 | 0.87 | 0.79 | 0.50 |
| 11 | 785.33 | 10.89 | 195.21 | 16.58 | 0.31 | 242.51 | 345.37 | 2113.29 | 102.55 | 11.46 | 4.58 | 19.88 | 40.46 | 0.84 | 0.64 | 0.87 | 0.79 | 0.50 |
| 12 | 792.32 | 7.83 | 204.33 | 16.74 | 0.30 | 246.62 | 347.13 | 2159.77 | 105.23 | 10.06 | 4.14 | 19.09 | 40.66 | 0.84 | 0.64 | 0.87 | 0.78 | 0.50 |
| 13 | 795.56 | 5.03 | 212.60 | 16.93 | 0.28 | 249.84 | 346.87 | 2199.74 | 107.08 | 9.16 | 3.79 | 18.40 | 41.07 | 0.84 | 0.63 | 0.87 | 0.78 | 0.50 |
| 14 | 806.46 | 2.21 | 220.63 | 17.21 | 0.27 | 253.53 | 348.14 | 2251.63 | 109.64 | 7.97 | 3.40 | 17.55 | 41.43 | 0.84 | 0.63 | 0.87 | 0.77 | 0.50 |



| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | COR | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| EUC | 1.00 | -0.15 | 0.30 | 0.40 | -0.20 | 0.23 | 0.45 | 0.53 | 0.47 | -0.16 | 0.10 | -0.10 | -0.12 | -0.15 | 0.15 | -0.09 | -0.28 |
| MAN | -0.15 | 1.00 | -0.34 | -0.17 | 0.32 | -0.36 | -0.26 | -0.26 | -0.13 | 0.23 | -0.07 | 0.05 | 0.20 | 0.12 | 0.04 | 0.19 | 0.23 |
| EYE | 0.30 | -0.34 | 1.00 | 0.37 | -0.35 | 0.78 | 0.45 | 0.57 | 0.46 | -0.28 | 0.27 | -0.21 | -0.20 | -0.04 | 0.07 | -0.08 | -0.19 |
| LEV | 0.40 | -0.17 | 0.37 | 1.00 | -0.27 | 0.30 | 0.34 | 0.65 | 0.24 | -0.24 | 0.09 | -0.10 | -0.17 | 0.04 | 0.15 | 0.03 | -0.09 |
| SMT | -0.20 | 0.32 | -0.35 | -0.27 | 1.00 | -0.30 | -0.26 | -0.30 | -0.40 | 0.26 | -0.23 | 0.26 | 0.10 | 0.15 | 0.08 | 0.12 | 0.29 |
| HAU | 0.23 | -0.36 | 0.78 | 0.30 | -0.30 | 1.00 | 0.45 | 0.48 | 0.32 | -0.23 | 0.21 | -0.11 | -0.18 | -0.05 | 0.03 | -0.10 | -0.18 |
| FRE | 0.45 | -0.26 | 0.45 | 0.34 | -0.26 | 0.45 | 1.00 | 0.75 | 0.42 | -0.13 | 0.13 | -0.07 | -0.05 | -0.13 | 0.12 | -0.10 | -0.25 |
| DTW | 0.53 | -0.26 | 0.57 | 0.65 | -0.30 | 0.48 | 0.75 | 1.00 | 0.47 | -0.23 | 0.13 | -0.13 | -0.12 | -0.03 | 0.18 | -0.05 | -0.20 |
| TDE | 0.47 | -0.13 | 0.46 | 0.24 | -0.40 | 0.32 | 0.42 | 0.47 | 1.00 | -0.18 | 0.13 | -0.26 | -0.02 | -0.07 | 0.13 | -0.05 | -0.20 |
| REC | -0.16 | 0.23 | -0.28 | -0.24 | 0.26 | -0.23 | -0.13 | -0.23 | -0.18 | 1.00 | 0.24 | 0.18 | 0.62 | -0.01 | -0.05 | 0.01 | 0.09 |
| DET | 0.10 | -0.07 | 0.27 | 0.09 | -0.23 | 0.21 | 0.13 | 0.13 | 0.13 | 0.24 | 1.00 | 0.06 | 0.05 | -0.07 | -0.04 | -0.03 | -0.14 |
| LAM | -0.10 | 0.05 | -0.21 | -0.10 | 0.26 | -0.11 | -0.07 | -0.13 | -0.26 | 0.18 | 0.06 | 1.00 | 0.01 | -0.06 | -0.05 | -0.01 | 0.02 |
| COR | -0.12 | 0.20 | -0.20 | -0.17 | 0.10 | -0.18 | -0.05 | -0.12 | -0.02 | 0.62 | 0.05 | 0.01 | 1.00 | 0.01 | -0.03 | 0.02 | 0.08 |
| MM_S | -0.15 | 0.12 | -0.04 | 0.04 | 0.15 | -0.05 | -0.13 | -0.03 | -0.07 | -0.01 | -0.07 | -0.06 | 0.01 | 1.00 | 0.50 | 0.52 | 0.53 |
| MM_A | 0.15 | 0.04 | 0.07 | 0.15 | 0.08 | 0.03 | 0.12 | 0.18 | 0.13 | -0.05 | -0.04 | -0.05 | -0.03 | 0.50 | 1.00 | 0.19 | 0.26 |
| MM_L | -0.09 | 0.19 | -0.08 | 0.03 | 0.12 | -0.10 | -0.10 | -0.05 | -0.05 | 0.01 | -0.03 | -0.01 | 0.02 | 0.52 | 0.19 | 1.00 | 0.36 |
| MM_P | -0.28 | 0.23 | -0.19 | -0.09 | 0.29 | -0.18 | -0.25 | -0.20 | -0.20 | 0.09 | -0.14 | 0.02 | 0.08 | 0.53 | 0.26 | 0.36 | 1.00 |

Table A.1: The Spearman rank correlation among different metrics across algorithms (SPP)

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AIM | 593.92 | 22.19 | 116.54 | 11.61 | 0.58 | 158.59 | 273.56 | 1444.21 | 86.60 | 0.35 | 0.00 | 0.20 | 19.86 | 0.77 | 0.42 | 0.76 | 0.66 |
| AWS | 593.23 | 23.72 | 109.86 | 11.38 | 0.58 | 161.36 | 271.98 | 1407.45 | 85.77 | 0.57 | 0.00 | 0.43 | 18.85 | 0.76 | 0.41 | 0.75 | 0.66 |
| CAS | 532.63 | 21.02 | 112.93 | 11.50 | 0.58 | 152.50 | 242.07 | 1312.11 | 75.73 | 0.44 | 0.00 | 0.16 | 20.11 | 0.78 | 0.41 | 0.78 | 0.68 |
| CVS | 659.01 | 52.73 | 170.68 | 12.43 | 0.57 | 213.22 | 304.00 | 1735.46 | 100.06 | 0.16 | 0.00 | 0.06 | 22.62 | 0.77 | 0.43 | 0.75 | 0.65 |
| DVA | 615.52 | 27.60 | 121.59 | 11.84 | 0.58 | 166.80 | 280.03 | 1491.45 | 89.73 | 0.35 | 0.00 | 0.26 | 18.96 | 0.76 | 0.42 | 0.75 | 0.65 |
| GBVS | 487.70 | 7.35 | 100.39 | 11.13 | 0.58 | 130.77 | 215.79 | 1176.68 | 69.77 | 0.74 | 0.00 | 0.33 | 19.59 | 0.79 | 0.42 | 0.79 | 0.70 |
| IKN | 550.60 | 16.57 | 114.01 | 11.49 | 0.58 | 151.81 | 250.14 | 1347.33 | 80.05 | 0.43 | 0.00 | 0.21 | 19.38 | 0.77 | 0.42 | 0.77 | 0.67 |
| IMSIG | 562.50 | 23.95 | 110.11 | 11.38 | 0.58 | 152.12 | 258.72 | 1346.12 | 77.43 | 0.60 | 0.00 | 0.28 | 19.04 | 0.77 | 0.41 | 0.76 | 0.66 |
| QSS | 672.74 | 40.35 | 143.32 | 12.11 | 0.57 | 203.34 | 312.80 | 1675.07 | 103.47 | 0.27 | 0.00 | 0.13 | 19.73 | 0.76 | 0.42 | 0.75 | 0.65 |
| SSR | 639.80 | 32.53 | 132.01 | 11.83 | 0.58 | 176.81 | 295.52 | 1558.77 | 92.35 | 0.41 | 0.00 | 0.17 | 19.93 | 0.77 | 0.42 | 0.76 | 0.66 |
| SUN | 613.25 | 30.00 | 128.63 | 11.89 | 0.58 | 172.51 | 281.01 | 1506.95 | 90.78 | 0.28 | 0.00 | 0.22 | 19.68 | 0.77 | 0.43 | 0.75 | 0.65 |
| cG | 425.96 | 25.20 | 124.23 | 11.30 | 0.59 | 153.05 | 186.99 | 1147.19 | 60.49 | 0.39 | 0.00 | 0.10 | 22.07 | 0.77 | 0.21 | 0.70 | 0.73 |
| SAM-VGG | 502.56 | 6.08 | 66.11 | 10.20 | 0.58 | 96.95 | 227.88 | 1052.29 | 68.58 | 0.64 | 0.00 | 1.07 | 16.90 | 0.78 | 0.43 | 0.77 | 0.70 |
| opensalicon | 493.05 | 2.53 | 75.73 | 10.35 | 0.58 | 105.61 | 227.40 | 1064.80 | 66.15 | 0.81 | 0.00 | 0.72 | 16.81 | 0.77 | 0.41 | 0.76 | 0.68 |
| salgan | 815.41 | 69.93 | 158.61 | 11.96 | 0.59 | 337.08 | 436.59 | 2054.97 | 144.33 | 0.50 | 0.00 | 0.78 | 16.77 | 0.75 | 0.43 | 0.73 | 0.66 |
| pathgan | 878.71 | 165.25 | 330.67 | 13.88 | 0.55 | 327.80 | 408.63 | 2598.86 | 128.85 | 0.00 | 0.06 | 1.42 | 19.95 | 0.57 | 0.21 | 0.48 | 0.46 |

Table A.2: The benchmark - SPP results

| Time | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 600 | 216.10 | 52.81 | 166.30 | 4.18 | 0.42 | 188.70 | 217.59 | 420.02 | 26.95 | 24.41 | 0.97 | 13.87 | 80.40 | 0.89 | 0.57 | 0.89 | 0.88 |
| 800 | 290.85 | 49.15 | 155.68 | 5.50 | 0.42 | 203.50 | 248.20 | 575.89 | 65.88 | 23.06 | 2.33 | 17.28 | 62.87 | 0.87 | 0.56 | 0.88 | 0.86 |
| 1000 | 358.09 | 46.80 | 144.83 | 6.81 | 0.43 | 209.60 | 268.46 | 726.90 | 82.44 | 21.68 | 3.35 | 18.97 | 53.28 | 0.86 | 0.56 | 0.88 | 0.85 |
| 1200 | 422.14 | 45.17 | 135.30 | 8.10 | 0.43 | 212.52 | 285.75 | 874.13 | 85.51 | 20.20 | 4.07 | 19.85 | 47.69 | 0.86 | 0.58 | 0.88 | 0.84 |
| 1400 | 480.79 | 44.11 | 127.39 | 9.32 | 0.43 | 212.85 | 299.27 | 1019.08 | 84.85 | 19.01 | 4.61 | 20.53 | 44.33 | 0.85 | 0.60 | 0.88 | 0.84 |
| 1600 | 528.57 | 43.72 | 121.09 | 10.35 | 0.43 | 211.13 | 307.42 | 1137.91 | 83.39 | 18.05 | 5.02 | 21.01 | 42.37 | 0.86 | 0.62 | 0.89 | 0.83 |
| 1800 | 559.07 | 43.52 | 117.55 | 11.01 | 0.43 | 209.98 | 312.34 | 1214.99 | 82.16 | 17.45 | 5.24 | 21.26 | 41.31 | 0.86 | 0.63 | 0.89 | 0.83 |
| 2000 | 569.69 | 43.45 | 116.50 | 11.20 | 0.43 | 209.56 | 313.70 | 1236.92 | 81.83 | 17.21 | 5.30 | 21.31 | 40.99 | 0.86 | 0.64 | 0.89 | 0.83 |
| 2200 | 570.96 | 43.49 | 116.40 | 11.23 | 0.43 | 209.51 | 313.91 | 1239.15 | 81.79 | 17.18 | 5.31 | 21.31 | 40.94 | 0.86 | 0.64 | 0.89 | 0.83 |
| 2400 | 571.02 | 43.49 | 116.40 | 11.23 | 0.43 | 209.51 | 313.89 | 1239.25 | 81.79 | 17.18 | 5.31 | 21.31 | 40.94 | 0.86 | 0.64 | 0.89 | 0.83 |
| 2600 | 572.29 | 43.23 | 118.23 | 11.05 | 0.43 | 211.45 | 315.94 | 1230.42 | 83.50 | 17.22 | 5.26 | 20.75 | 41.47 | 0.86 | 0.63 | 0.88 | 0.83 |
| 2800 | 703.85 | 44.50 | 101.73 | 13.97 | 0.45 | 196.94 | 326.33 | 1559.30 | 76.91 | 24.34 | 8.02 | 26.48 | 39.39 | 0.86 | 0.69 | 0.89 | 0.83 |

Table A.3: Saliency in time - SPP results

| | EUC | MAN | EYE | LEV | SMT | HAU | FRE | DTW | TDE | REC | DET | LAM | CORM | MM_S | MM_A | MM_L | MM_P | MM_D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Affective | 685.96 | 32.24 | 107.11 | 9.30 | 0.14 | 190.35 | 201.39 | 2712.41 | 53.60 | 6.11 | 5.57 | 9.82 | 17.03 | 0.04 | 0.17 | 0.06 | 0.06 | 0.19 |
| BlackWhite | 659.40 | 29.71 | 100.18 | 9.12 | 0.14 | 180.38 | 192.49 | 2524.98 | 52.31 | 5.90 | 5.53 | 10.07 | 18.35 | 0.04 | 0.16 | 0.06 | 0.06 | 0.19 |
| Fractal | 761.61 | 31.14 | 100.95 | 9.13 | 0.13 | 183.53 | 197.63 | 2613.20 | 56.01 | 6.08 | 5.47 | 10.30 | 19.26 | 0.04 | 0.16 | 0.06 | 0.06 | 0.18 |
| Inverted | 686.92 | 30.29 | 96.57 | 8.58 | 0.13 | 172.23 | 187.76 | 2431.32 | 52.90 | 4.75 | 5.31 | 10.08 | 19.64 | 0.03 | 0.14 | 0.05 | 0.06 | 0.18 |
| Line Drawing | 708.23 | 29.50 | 93.04 | 8.78 | 0.12 | 168.59 | 177.14 | 2491.01 | 54.75 | 6.06 | 5.26 | 10.37 | 20.35 | 0.03 | 0.14 | 0.05 | 0.06 | 0.18 |
| Noisy | 622.31 | 30.75 | 100.73 | 9.89 | 0.15 | 178.69 | 197.67 | 2421.20 | 58.80 | 8.56 | 5.64 | 10.92 | 20.08 | 0.04 | 0.18 | 0.06 | 0.06 | 0.20 |
| Outdoor | 683.76 | 30.95 | 96.43 | 9.06 | 0.14 | 184.42 | 198.42 | 2465.37 | 52.76 | 6.48 | 5.49 | 10.13 | 18.84 | 0.04 | 0.17 | 0.06 | 0.06 | 0.19 |
| Pattern | 653.08 | 31.71 | 105.42 | 9.80 | 0.14 | 180.26 | 195.58 | 2639.35 | 56.26 | 7.19 | 5.43 | 10.55 | 19.83 | 0.04 | 0.16 | 0.06 | 0.06 | 0.19 |
| Satellite | 719.24 | 32.91 | 106.05 | 8.91 | 0.14 | 185.62 | 203.48 | 2628.73 | 58.27 | 5.88 | 5.42 | 10.58 | 19.64 | 0.04 | 0.16 | 0.06 | 0.06 | 0.19 |
| Social | 712.32 | 30.71 | 96.41 | 9.39 | 0.14 | 180.79 | 205.10 | 2691.31 | 49.06 | 6.68 | 5.71 | 9.85 | 17.33 | 0.04 | 0.18 | 0.05 | 0.06 | 0.19 |

Table A.4: Standard deviation of Inter observer distances per category for CAT2000

# Bibliography

[1] American academy of ophthalmology. https://www.aao.org/, Sept. 2018. v, 10

[2] H. Adeli and G. Zelinsky. Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1932–1942, 2018. 3

[3] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015. 29, 32, 50

[4] N. C. Anderson, W. F. Bischof, K. E. Laidlaw, E. F. Risko, and A. Kingstone. Recurrence quantification analysis of eye movements. *Behavior research methods*, 45(3):842–856, 2013. 30

[5] N. C. Anderson, W. F. Bischof, K. E. W. Laidlaw, E. F. Risko, and A. Kingstone. Recurrence quantification analysis of eye movements. *Behavior Research Methods*, 45(3):842–856, Sep 2013. 21

[6] B. Aronov, S. Har-Peled, C. Knauer, Y. Wang, and C. Wenk. Fréchet distance for curves, revisited. In *European Symposium on Algorithms*, pages 52–63. Springer, 2006. 27

[7] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. OConnor. Saltinet: Scanpath prediction on 360 degree images using saliency volumes. In *ICCV Workshop*, 2017. 17

[8] D. J. Berndt. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994. 21, 26

[9] J. Biondi, G. Fernandez, S. Castro, and O. Agamennoni. Eye-movement behavior identification for ad diagnosis. *arXiv preprint arXiv:1702.00837*, 2017. 38

[10] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 13, 41

[11] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013. 33

[12] N. Bruce. Evolutionary design for computational visual attention. Master's thesis, University of Waterloo, 2003. 10

[13] N. Bruce, C. Catton, and S. Janjic. A deeper look at saliency: Feature contrast, semantics, and beyond. pages 516–524, 2016. 3

[14] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006. 16

[15] M. Burmester and M. Mast. Repeated web page visits and the scanpath theory: A recurrent pattern detection approach. *Journal of Eye Movement Research*, 3(4), 2010. 17

[16] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 13

[17] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. 12

[18] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 2018. vii, 19, 20

[19] Z. Chen and W. Sun. Scanpath prediction for visual attention using ior-roi lstm. In *IJCAI*, pages 642–648, 2018. 18

[20] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010. 20

[21] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010. 21, 23

[22] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist. It depends on how you look at it: Scanpath comparison in mul-

tiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012. v, 20, 21, 39, 46

[23] A. T. Duchowski, J. Driver, S. Jolaoso, W. Tan, B. N. Ramey, and A. Robbins. Scanpath comparison revisited. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 219–226. ACM, 2010. 20

[24] T. Eiter and H. Mannila. Computing discrete fréchet distance. Technical report, Citeseer, 1994. 21

[25] T. Eiter and H. Mannila. Computing discrete fréchet distance. Technical report, Citeseer, 1994. 27

[26] M. Emami and L. L. Hoberock. Selection of a best metric and evaluation of bottom-up visual saliency models. *Image and Vision Computing*, 31(10):796–808, 2013. 19

[27] T. Foulsham, R. Dewhurst, M. Nyström, H. Jarodzka, R. Johansson, G. Underwood, and K. Holmqvist. Comparing scanpaths during scene encoding and recognition: A -dimensional approach. 2012. 17

[28] T. Foulsham, R. Dewhurst, M. Nyström, H. Jarodzka, R. Johansson, G. Underwood, and K. Holmqvist. Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Journal of Eye Movement Research*, 5(4), 2012. 23

[29] T. Foulsham and G. Underwood. What can saliency models predict about eye

movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of vision*, 8(2):6–6, 2008. 17, 26

[30] J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on robotics and automation*, 16(6):890–898, 2000. 17

[31] J. Harold, I. Lorenzoni, T. F. Shipley, and K. R. Coventry. Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change*, 6(12):1080, 2016. 17

[32] J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye movements*, pages 537–III. Elsevier, 2007. 26

[33] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 16

[34] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993. 21, 28

[35] M. A. Islam, M. Kalash, and N. D. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *In Proceedings of the IEEE International Conference on Computer Vision*, 2018. 3

[36] M. A. Islam, M. Kalash, M. Rochan, N. D. Bruce, and Y. Wang. Salient object

detection using a context-aware refinement network. In *British Machine Vision Conference*, 2017. 3

[37] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 16, 17, 58

[38] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1072–1080. IEEE, 2015. 13

[39] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 1072–1080. IEEE, 2015. 16, 64

[40] M. Jiang, J. Xu, and Q. Zhao. Saliency in crowd. In *European Conference on Computer Vision*, pages 17–32. Springer, 2014. 13, 41

[41] T. Judd, F. Durand, and A. Torralba. Fixations on low resolution images. *Journal of Vision*, 10(7):142–142, 2010. 62

[42] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 33

[43] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1985. 3, 16, 17

[44] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein. What do saliency models predict? *Journal of vision*, 14(3):14–14, 2014. 16

[45] I. Kotseruba, A. Rasouli, and J. K. Tsotsos. Joint attention in autonomous driving (jaad). *arXiv preprint arXiv:1609.04741*, 2016. 17

[46] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 16

[47] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 16

[48] W. Kuo, B. Hariharan, and J. Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2479–2487, 2015. 3

[49] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013. 19

[50] O. Le Meur, T. Baccino, and A. Roumy. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 373–382. ACM, 2011. 33

[51] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen. A data-driven metric for compre-

hensive evaluation of saliency models. In *Proceedings of the IEEE international conference on computer vision*, pages 190–198, 2015. 19

[52] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 13

[53] S. K. Mannan, K. H. Ruddock, and D. S. Wooding. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial vision*, 10(3):165–188, 1996. 21, 25

[54] S. Mathôt, F. Cristino, I. D. Gilchrist, and J. Theeuwes. A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research*, 5(1), 2012. 21, 26, 37, 39

[55] N. Mavridis. *Grounded situation models for situated conversational assistants.* PhD thesis, Massachusetts Institute of Technology, 2007. 17

[56] L. Meur. Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164, 2015. 17, 29

[57] K. Nakayama and M. Mackeben. Sustained and transient components of focal visual attention. *Vision research*, 29(11):1631–1647, 1989. 9

[58] V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005. 16

[59] T. Ngo and B. Manjunath. Saccade gaze prediction using a recurrent neural network. 2017. 18

[60] D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971. 17, 22

[61] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 16

[62] E. Pellicano, A. D. Smith, F. Cristino, B. M. Hood, J. Briscoe, and I. D. Gilchrist. Children with autism are neither systematic nor optimal foragers. *Proceedings of the National Academy of Sciences*, 108(1):421–426, 2011. 23

[63] C. M. Privitera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on pattern analysis and machine intelligence*, 22(9):970–982, 2000. 21, 22

[64] S. Rahman and N. D. Bruce. Factors underlying inter-observer agreement in gaze patterns: Predictive modelling and analysis. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 155–162. ACM, 2016. 33

[65] A. Rasouli and J. K. Tsotsos. Visual saliency improves autonomous visual search. In *Computer and Robot Vision (CRV), 2014 Canadian Conference on*, pages 111–118. IEEE, 2014. 17

[66] A. G. Samuel and D. Kat. Inhibition of return: A graphical meta-analysis of

its time course and an empirical test of its temporal and spatial properties. *Psychonomic Bulletin & Review*, 10(4):897–906, 2003. 2

[67] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991. 28

[68] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of vision*, 9(12):15–15, 2009. 16

[69] Z. Sharafi, Z. Soh, and Y.-G. Guéhéneuc. A systematic literature review on the usage of eye-tracking in software engineering. *Information and Software Technology*, 67:79–107, 2015. 23

[70] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 16

[71] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659, 2005. 26

[72] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 16

[73] J. K. Tsotsos, L. Itti, and G. Rees. A brief and selective history of attention. 2005. 12

[74] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. In *Computer Vision and Pattern*

*Recognition (CVPR), 2011 IEEE Conference on*, pages 441–448. IEEE, 2011. 21, 28

[75] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 13, 41

[76] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 3

[77] A. L. Yarbus. Eye movement and vision, trans. b. haigh. *ed: Plenum Press, New York*, 1967. 10

[78] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7):32–32, 2008. 16

[79] D. Zhao, Y. Ma, J. Wang, and Z. Jiang. Hierarchical reinforcement learning for saliency detection of low-resolution airports. In *IGARSS*, pages 1622–1625, 2016. 16