# Bayesian inference for an exploration-exploitation model of human gaze control

**Noa Malem-Shinitski (malem@uni-potsdam.de)**
Institute of Mathematics, Potsdam University, Karl-Liebknecht-Str. 24-25
Potsdam, 14476 Germany

**Stefan Seelig**
Department of Psychology, Potsdam University, Karl-Liebknecht-Str. 24-25
Potsdam, 14476 Germany

**Sebastian Reich**
Institute of Mathematics, Potsdam University, Karl-Liebknecht-Str. 24-25
Potsdam, 14476 Germany

**Ralf Engbert**
Department of Psychology, Potsdam University, Karl-Liebknecht-Str. 24-25
Potsdam, 14476 Germany

## Abstract

**Understanding human gaze, and the saccadic selection process underlying it, is an important question in cognitive-neuroscience with many interesting applications in areas from psychology to computer vision. One way to advance our understanding is to develop generative models that capture the spatial interaction between fixations and the temporal structure of a sequence of fixations, known as scanpaths. Such models are scarce in the literature and even fewer attempt to model inter-subject variability. In this work, we present a new parametric model for scanpath generation. We develop a discrete-time probabilistic generative model, with a Markovian structure, where at each step the next fixation location is selected using one of two strategies - exploitation or exploration. We implement efficient Bayesian inference for hyperparameter estimation using an HMC within Gibbs approach. Our model is able to capture inter-observer variability in terms of saccade length and direction as demonstrated by fitting the model to a dataset of scanpaths from 35 subjects performing a task of free viewing of 30 natural scene image.**

**Keywords:** generative model, eye movement, attention, bayesian inference, scanpath generation

## Introduction

In an attempt to understand the underlying cognitive mechanisms of human vision, much work has been done trying to answer the question - what do human observers look at in an image? To answer this question many saliency models were developed (Itti & Koch, 2001; Kümmerer, Wallis, Gatys, & Bethge, 2017), which generate fixation density estimates which predict the density of human fixations in an image.

Not only is it important to quantify where human observers look in an image, but it is also important to quantify how they look at an image. Rapid eye movements are performed between two fixation points; these movements are referred to as saccades. A sequence of saccades is called a scanpath. Understanding how human observers look at an image requires quantifying scanpaths.

Several models for scanpath generation and prediction were developed in recent years. Earlier models rely on cognitive and neural assumptions regarding human perception (Le Meur & Liu, 2015; Engbert, Trukenbrod, Barthelmé, & Wichmann, 2015) and with the rise of machine learning, and deep learning in particular, several models have been developed which employ state of the art deep learning techniques (Shao et al., 2017; Kmmerer, Wallis, & Bethge, 2018). While these models are very successful in capturing the properties of scanpaths across a lare group of human observers, they require a lot of data for training and cannot be fitted for individual observers.

In this work we present a stochastic generative model which generates a scanpath given a saliency map. As the model is relatively simple and includes only very few parameters, we can fit it to data from individual experimental subjects and capture inter-subject variability.

Next we describe the model in details, the assumptions behind it and the inference procedure used to fit it to experimental data. Then we present the result of the fitted model on test data and conclude by discussing the limitations of the model and future research directions.

## Methods

### Exploration-Exploitation Model

Our model is Markovian and we assume that conditioned on the current fixation location the next fixation location, is independent of the rest of the scanpath. The novelty of our work is that we use the well known concept of exploration and exploitation to generate the next fixation. We use this concept to characterize the spatial properties of saccades generation. This approach is motivated by previous research that used the Exploration-Exploitation framework to characterize the temporal structure of saccade generation (Gameiro, Kaspar, König,

Nordholt, & König, 2017).

Since our main focus is on modeling the dynamics of the saccade generation mechanism, we assume that we have a "good enough" model for saliency maps generation. In practice, for each image we use the empirical saliency over all the participants similarly to (Schütt et al., 2017). A saliency map is nothing but a function $\mu(z) : \mathbb{R}^2 \mapsto \mathbb{R}^+$ with $z = (x, y)$ being a location in an image and $\mu(z)$ the probability of an average viewer to fixate on this location. From now on we will use $\mu(z)$ to refer to the saliency of the image in $z$.

Generally, scanpaths are sequences of fixation locations and duration. In this work we model only the spatial properties of gaze control and do not model the temporal dimension. Thus, a scanpath can be written as $Z = \{z_1, z_2, ..., z_t, ..., z_T\}$ with $T$ the number of fixations in the scanpath and $z_t$ the location of the $t$th fixation.

In our model, for every time $t$ given the current fixation location $z_{t-1}$ the next fixation $z_t$ is generated following one of two policies:

**Exploitation** Given that the current fixation location is "interesting enough" (e.g. the value of $\mu(z_{t-1})$ is high) the "viewer" assumes that there are further interesting locations near by and the next fixation is generated as a small brownian step around the current location with variance $\varepsilon$. This is written as:

$$p(z_t|z_{t-1}) = n(z_t; z_{t-1}, \varepsilon) \tag{1}$$

where $n(z_t; z_{t-1}, \varepsilon)$ is a Gaussian density with mean $z_{t-1}$ and variance $\varepsilon$.

**Exploration** On the other hand, the viewer may decide that the current fixation location is not interesting enough. In this case the viewer would chose randomly the next fixation according to the saliency map. This policy may lead to very large saccade amplitudes which are known to be less probable (Tatler, Baddeley, & Vincent, 2006). To encapsulate this prior knowledge we perform a point-wise multiplication of the saliency map with a Gaussian distribution. This multiplication does not necessarily result in a valid density and it requires normalization, leading to the following expression:

$$p(z_t|z_{t-1}) = \frac{\mu(z_t)\, n(z_t; z_{t-1}, \xi)}{\sum_{z'} \mu(z')\, n(z'; z_{t-1}, \xi)}. \tag{2}$$

Figure 1 visualizes the two distributions formulated in Equations 1 and 2.

At each step the choice of the strategy is made by sampling from a Bernouli distribution, written as:

$$p(z_t|z_{t-1}) = (n(z_t; z_{t-1}, \varepsilon))^{\gamma_t} \left( \frac{\mu(z_t)\, n(z_t; z_{t-1}, \xi)}{\sum_{z'} \mu(z')\, n(z'; z_{t-1}, \xi)} \right)^{1-\gamma_t} \tag{3}$$

$$p(\gamma_t) = Bern(\gamma_t; \rho). \tag{4}$$

Our next assumption is that the decision whether to make an exploration or an exploitation step depends on the saliency value of the current fixated location. The result is that the viewer is more likely to make a small exploitation step if the current fixated location is already of high interest. We capture this in the model by allowing the bias of the parameter $\gamma_t$ to be dependent on the saliency in $z_{t-1}$, specifically:

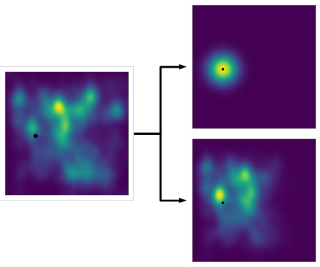

Figure 1: On the left is an example for a saliency map (brighter colors represent areas of higher interest) and a fixation location. On the right are the two distributions from which the next fixation location may be drawn according to the two different policies. The upper tile corresponds to the Exploitation policy and to Equation 1. The lower tile corresponds to the Exploration policy and to Equation 2.

$$p(\gamma_t|z_{t-1}) = Bern(\gamma_t; \rho_{t-1}) \tag{5}$$

$$\rho_{t-1} = \sigma(\mu(z_{t-1})) = \frac{1}{1 + \exp(-b(\mu(z_{t-1}) - \mu_0))}. \tag{6}$$

We chose the Sigmoidal link function to induce smoothness and to simplify the inference process.

The generation of a scanpath in our model is sequential. At each time step, the next fixation location is generated by sampling $\gamma_t$ form the Bernouli distribution in Equation 5. If $\gamma_t = 1$ the next fixation location $z_t$ is sampled from the density in Equation 1 and if $\gamma_t = 0$, from Equation 2. In this framework the vector $\Gamma = \{\gamma_1, ..., \gamma_T\}$ can be seen as unobserved data and we can write the likelihood of the partially observed data as:

$$p(Z, \Gamma|\Theta) \approx p(Z|\Gamma, \Theta)\, p(\Gamma|\Theta) = p(z_1) \prod_{t=2}^{t=T} p(z_t|z_{t-1})\, \rho_{t-1}^{\gamma_t} (1 - \rho_{t-1})^{(1-\gamma_t)} \tag{7}$$

$$\Theta = \{b, \mu_0, \varepsilon, \xi\} \tag{8}$$

## Inference

We want to have a unique generative model for each subject. Thus we fit the model parameters $\Theta$ separately for the data from each subject in a Bayesian framework. This approach allows us to include prior knowledge regarding the different model parameters based on known spatial features of scan paths.

Naively one could chose prior distributions over the model parameters and maximize the posterior. As the posterior of our model cannot be maximized analytically we resort to sampling. We implement a method known as MCMC within Gibbs

sampler (Gilks & Wild, 1992) where an Hamiltonian Monte Carlo sampler is used to sample from the conditionals of the parameters for which there is no conjugate structure.

## Results

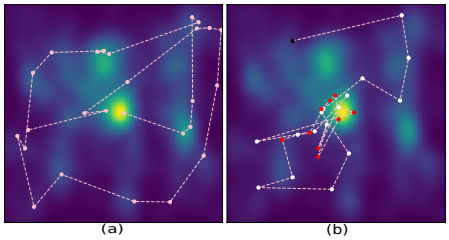


Figure 2: (a) An empirical scanpath and the corresponding saliency map of the image that was viewed by an experimental subject. (b) A scanpath generated from the model fitted for the same subject. In red are fixation locations which are results of an exploitation step and in white are fixation locations that came after an exploration steps.

For validation of our model we used the dataset used in (Schütt et al., 2017). This data contains 35 viewers freely observing 30 different natural images.

Figure 2 (b) presents a scanpath generated by our model for a particular saliency map. Each dot presents a fixation location. The initial random fixation is shown in black. Fixation locations $z_t$ which are results of an exploitation step are shown in red, while those from exploration are shown in white. For comparison Figure 2 (a) presents the recorded scanpath of a subject observing the image from which the saliency map was produced. We present the saliency maps rather than the original images as the saliency maps are the input to the Exploration-Exploitation model and not the RGB images which are observed by the subjects in the experiment.

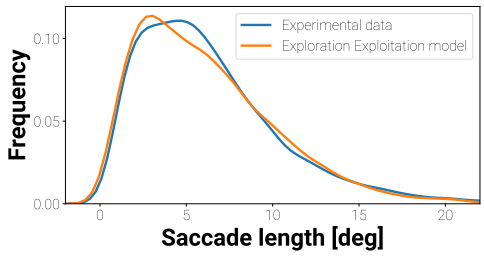


Figure 3: Frequency of saccade lengths (in visual degrees). In blue is the experimental test data and in orange data generated by the fitted model.

In order to assess the model, we compare the statistical properties of data generated by the model and empirical data across subjects. To do so, we followed a train-test framework. For each subject a model was fitted using data from $70\%$ of the images (21 images) and the comparison was made with respect to the remaining $30\%$ of the images (9 images) that were not used to infer the model parameters.

First, we look at the saccade length density of the data from all subjects, presented in Figure 3. Our model achieves very high agreement with the experimental test data. It seems though that the model generates more short saccades (around 3 visual degrees) and less medium length saccades (around 6 visual degrees) than the experimental subjects.

Not only do we compare the saccade length, but we also compare the saccade direction. Figure 4 presents the frequency of the saccade direction for the empirical data and the data generated from our model. Our model captures the tendency to perform horizontal saccades but fails to capture the bias towards vertical saccades. This is expected considering the Gaussian distributions that are used in the model.

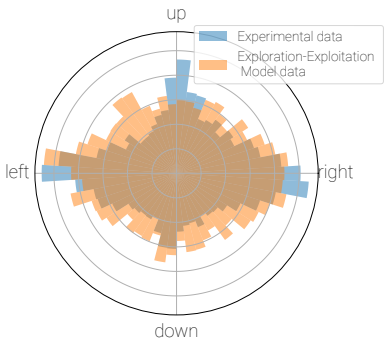


Figure 4: Saccade direction frequency. Blue - the experimental data. Orange - data generated by the fitted model. These results are for test data that was not used for fitting the model.

One of the main novelties of our model in comparison to existing models is its ability to capture inter-subject variability. To demonstrate this we compare in Figure 5 the median saccade length of each subject with the median saccade length from the data generated from a model fitted for that subject. The blue line is the identity line. The model does not capture perfectly the exact median saccade length for each subject but it does capture the variability between subjects and can potentially be used for subject identification.

## Discussion & Summary

In this work we presented a new model for scanpath generation which assumes that each saccade is generated following an exploration or exploitation policy. Not only is the model itself new, but our approach of Bayesian inference for fitting the model is rarely used in the field. A notable exception is the SceneWalk model (Schütt et al., 2017), where the Metropolis Hastings algorithm was used to infere the model parameters. To our knowledge there is no previous work on a model for scanpath generation which attempted to construct a likelihood with a conjugate structure and to use the Gibbs sampler for inference. Furthermore, we successfully fit the model to exper-

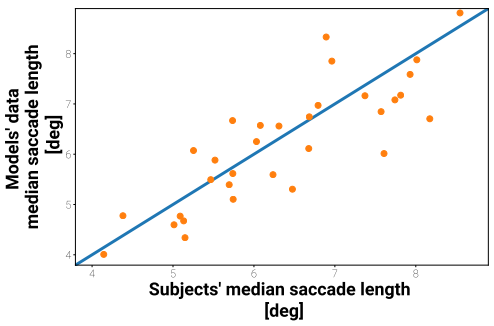


Figure 5: Comparison of the median saccade length of the experimental test data and data generated from the model. Each dot represents data from one subject. The blue line is the identity line.

imental data from individual subjects viewing only 30 different natural scenes images and capture the statistics of both the entire population and the individual subjects.

The exploration-exploitation model focuses on the dynamical aspect of a scanpath and we assume that the underlying image saliency is known. In practice we used the experimental saliency as the saliency input to the model. This approach is not valid for forecasting scanpaths over new images that were not observed by the subject. In these cases we would need to use a computational saliency model. Many saliency models are available and further work should be done to assess the effect of the choice of the computational saliency model on the performance of the exploration-exploitation model.

We presented the model performance in terms of capturing the experimental distribution of saccade directions. Another important aspect of scanpath generation is the angles between saccades. Examination of the empirical data showed that the angle between consecutive saccades is usually small, a phenomenon known as saccadic momentum (Smith & Henderson, 2009; Wilming, Harst, Schmidt, & König, 2013; Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2016). Currently, our model is not capable of capturing this phenomenon. This is expected since the likelihood has a Markov structure and includes information only of the location of the previous fixation but not the direction of the previous saccade. This may be solved by adding memory of fixations farther in the past to the likelihood and generating saccades in a Polar coordinate system rather than Cartesian.

Lastly, we would like to mention that as our model is relatively simple it can be used to implement different Bayesian inference algorithms and assess their performance in terms of computational time and accuracy. One such algorithm we would like to evaluate is Sequential Monte Carlo (Liu & Chen, 1998) which is often used in the field of data assimilation.

## Acknowledgments

This research has been funded by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1294 ”Data Assimilation”, Project B03 ”Parameter inference and model comparison in dynamical cognitive models”.

## References

Engbert, R., Trukenbrod, H. A., Barthelmé, S., & Wichmann, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *Journal of Vision*, *15*(1), 14–14.

Gameiro, R. R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and exploitation in natural viewing behavior. *Scientific Reports*, *7*(1), 2311.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *41*(2), 337–348.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*(1), 161–170.

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017, Oct). Understanding low- and high-level contributions to fixation prediction. In *The ieee international conference on computer vision (iccv)*.

Kmmerer, M., Wallis, T. S. A., & Bethge, M. (2018). Extending deepgaze ii: Scanpath prediction from deep features. In *Annual meeting of the vision sciences society (vss)*.

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Research*, *116*, 152–164.

Liu, J. S., & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American Statistical Association*, *93*(443), 1032–1044.

Rothkegel, L. O., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2016). Influence of initial fixation position in scene viewing. *Vision Research*, *129*, 33–49.

Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505.

Shao, X., Luo, Y., Zhu, D., Li, S., Itti, L., & Lu, J. (2017). Scanpath prediction based on high-level features and memory bias. In *International conference on neural information processing* (pp. 3–13).

Smith, T. J., & Henderson, J. M. (2009). Facilitation of return during scene viewing. *Visual Cognition*, *17*(6-7), 1083–1108.

Tatler, B. W., Baddeley, R. J., & Vincent, B. T. (2006). The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, *46*(12), 1857–1862.

Wilming, N., Harst, S., Schmidt, N., & König, P. (2013). Saccadic momentum and facilitation of return saccades contribute to an optimal foraging strategy. *PLoS Computational Biology*, *9*(1), e1002871.