

# Simulating Human Saccadic Scanpaths on Natural Images

<sup>1</sup>N E . <sup>3</sup> V U , C A S , 1 4 , C  
<sup>4</sup>D S M P S , P U , 1 1, C  
<sup>1,3</sup> C C <sup>1</sup>, <sup>1,2</sup> F F <sup>2,4</sup>, <sup>2</sup>,  
<sup>2</sup> M P M E , P U

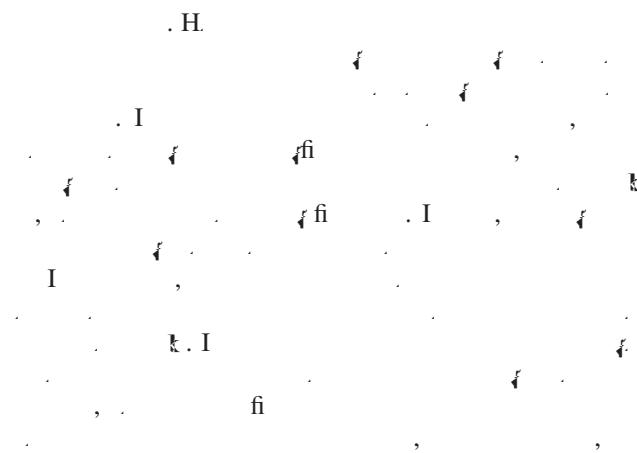
wwang@jdl.ac.cn, {chencheng880829, yizhou.wang, ttjiang, ffang, yuany}@pku.edu.cn

## Abstract

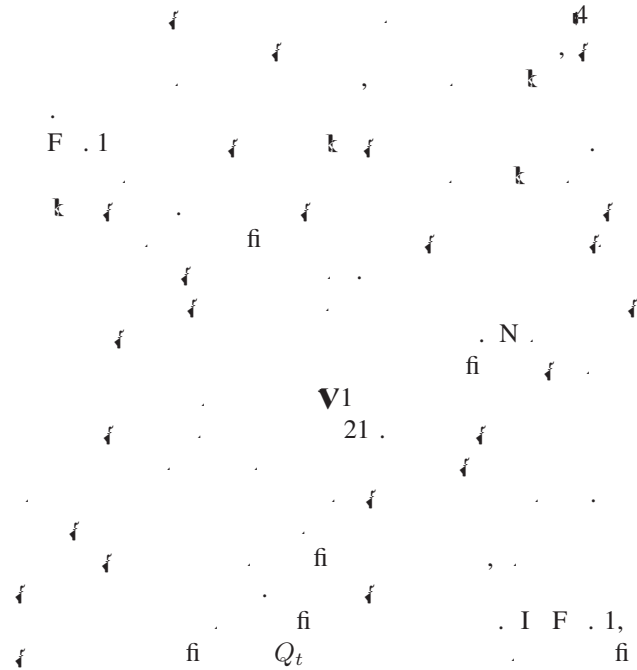
Human saccade is a dynamic process of information pursuit. Based on the principle of information maximization, we propose a computational model to simulate human saccadic scanpaths on natural images. The model integrates three related factors as driven forces to guide eye movements sequentially: reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory. For each eye movement, we compute three multi-band filter response maps as a coherent representation for the three factors. The three filter response maps are combined into multi-band residual filter response maps, on which we compute residual perceptual information (RPI) at each location. The RPI map is a dynamic saliency map varying along with eye movements. The next fixation is selected as the location with the maximal RPI value. On a natural image dataset, we compare the saccadic scanpaths generated by the proposed model and several other visual saliency-based models against human eye movement data. Experimental results demonstrate that the proposed model achieves the best prediction accuracy on both static fixation locations and dynamic scanpaths.

## 1. Introduction

In this paper, we propose a computational model to simulate human saccadic scanpaths on natural images. The model integrates three related factors as driven forces to guide eye movements sequentially: reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory. For each eye movement, we compute three multi-band filter response maps as a coherent representation for the three factors. The three filter response maps are combined into multi-band residual filter response maps, on which we compute residual perceptual information (RPI) at each location. The RPI map is a dynamic saliency map varying along with eye movements. The next fixation is selected as the location with the maximal RPI value. On a natural image dataset, we compare the saccadic scanpaths generated by the proposed model and several other visual saliency-based models against human eye movement data. Experimental results demonstrate that the proposed model achieves the best prediction accuracy on both static fixation locations and dynamic scanpaths.



### Proposed method





1. H, S, 3, dynamic, C, 1, R, et al. 22, N, H, et al. 14, et al. 1, F, et al. 2, 1, 2, S, 3, F, S, 4.

**2. Our Approach**

I, F, 1

**2.1. Coherent representation of three factors**

fi, F, 3, A, fi, O

**2.1.1 Sparse coding filters**

S, 3, V1, I, 21, multi-band filter response maps, C, A, ICA, 1, S, fi, I, 12, 8 x 8 x 3, F, 2, 4



F, 2, 4

**2.1.2 Foveal imaging**

P, 12, F, 11, S, A, F, 3



### 2.1.3 Visual working memory

$V$

#### Simulating the forgetting properties.

$I$

#### Updating visual working memory.

$V$

$A$  Max

$4$

$f_k^v(x, y, t)$

$f_k^w(x, y, t)$

$(x, y)$

$t$

$f_k^w(x, y, t) \leftarrow \max(f_k^v(x, y, t), \epsilon \cdot f_k^w(x, y, t - 1))$  1

#### Computing residual filter response maps.

$F$

$A$

$r_k = |f_k^o - f_k^w|$

$f_k^o$

$k$

### 2.2. Measuring residual perceptual information

$F$

$I$

$S_i$

$2$

$R$

$I$

$SER$

$SER$

$SER$

$$S_i = \sum_k SER_{ki} = - \sum_k (\pi_{ki} \sum_j P_{kij} \log P_{kij}) \quad 2$$

$\pi_{ki}$

$P_{kij}$

$A$

$2$

$SER$

$P$

$SER$

$SER$

$F$

$SER$

$M$

$e.g.$

$S$

$SER$

### 2.3. Saccadic amplitude

$U$

$90\%$

$20^\circ$

$Q_t$

$Q_{t+1}$

$Z \times Z$

$4$

$20^\circ$

$Z/2$

$SER$

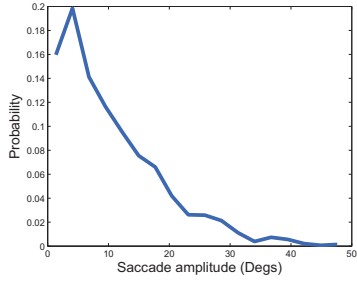
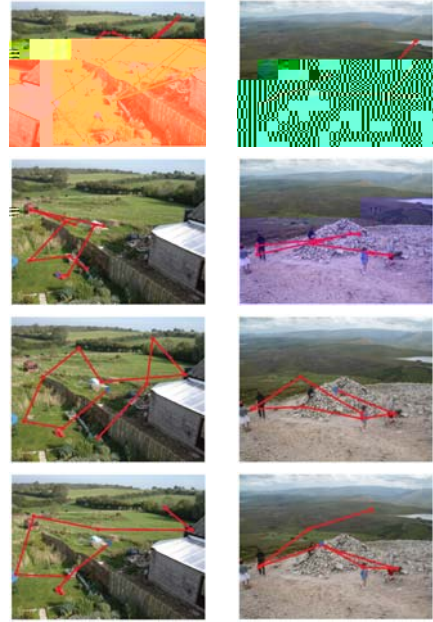


Figure 4.



$$Q_{t+1} \sim N(\mu, \sigma^2)$$

$$p(z \leq Z/2)$$

### 3. Experimental Results

Figure 5.

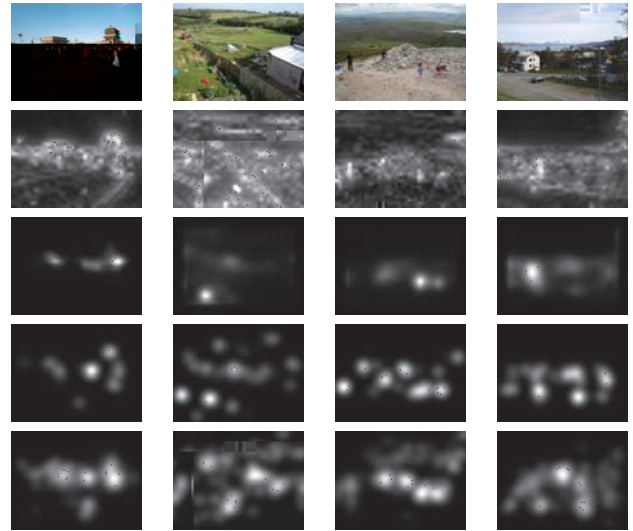
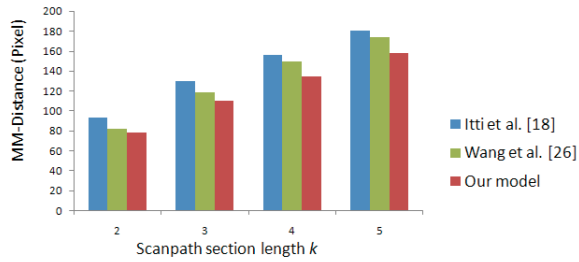
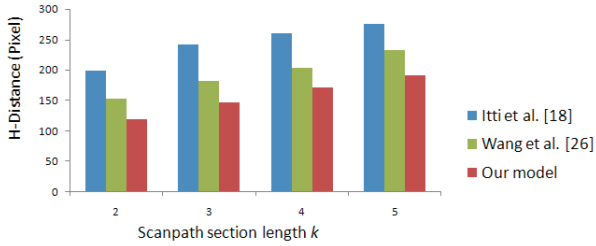
#### 3.1. Dataset and eye movement data collection

Our dataset consists of 24 scenes, each with 3 fixations. The scenes are categorized into 3 groups: 8 scenes with 21 fixations, 8 scenes with 21 fixations, and 8 scenes with 21 fixations. The scenes are collected from various sources, including YouTube and Flickr. The scenes are used to evaluate the performance of the proposed method.

#### 3.2. Evaluation of fixation order

We evaluate the performance of the proposed method by comparing it with the baseline method. The proposed method achieves a higher accuracy in predicting the fixation order compared to the baseline method. The results are shown in Table 1.

The proposed method is evaluated on a dataset of 24 scenes. The scenes are categorized into 3 groups: 8 scenes with 21 fixations, 8 scenes with 21 fixations, and 8 scenes with 21 fixations. The scenes are collected from various sources, including YouTube and Flickr. The scenes are used to evaluate the performance of the proposed method. The proposed method achieves a higher accuracy in predicting the fixation order compared to the baseline method. The results are shown in Table 1.



For each scanpath  $C$ , we generate a set of  $k$  scanpath sections  $\{C_m^k(t)\}_t$  of length  $k$ .

### 3.2.1 Distance of scanpaths

In this section, we define the distance between two scanpaths  $C$  and  $H$  based on the time-delay embedding  $C_m^k(t) = (c_m(t), \dots, c_m(t+k-1))$  of the scanpath  $C$ . The set of all possible scanpath sections of length  $k$  is denoted by  $X = \{C_m^k(t)\}_t \subseteq \mathbb{R}^k$ . Similarly, the set of all possible scanpath sections of length  $k$  for scanpath  $H$  is denoted by  $Y = \{C_h^k(\tau)\}_\tau \subseteq \mathbb{R}^k$ .

For a scanpath section  $x = C_m^k(t) \in X$ , the distance  $d_k(x, Y)$  between  $x$  and the set  $Y$  is defined as:

$$d_k(x, Y) = \min_{\tau} \{\|x - C_h^k(\tau)\|_2\} / k$$

The distance between two scanpaths  $C$  and  $H$  is defined as:

For a scanpath section  $x = C_m^k(t) \in X$ , the distance  $d_k(x, Y)$  between  $x$  and the set  $Y$  is defined as:

$$d_H^k = \max_t \{\min_{\tau} \{\|C_m^k(t) - C_h^k(\tau)\|_2\} / k\}$$

$$d_H^k = \max_t \{d_k(C_m^k(t), Y)\} \quad (3)$$

The mean minimal distance (MMD) between two scanpaths  $C$  and  $H$  is defined as:

$$d_M^k = E_t [d_k(C_m^k(t), Y)] \quad (4)$$

In our experiments, we set  $\epsilon = 0.7$ ,  $Z = 800$ , and  $\theta = 2.3^\circ$ .

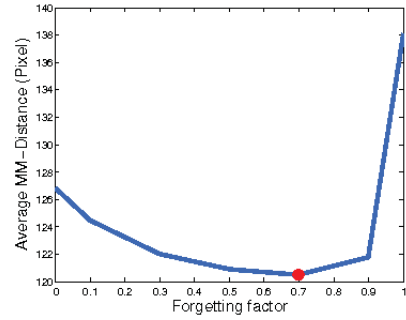
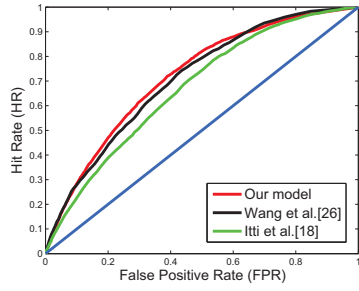


Fig. 1. ROC curves for different models. The x-axis represents the False Positive Rate (FPR) and the y-axis represents the Hit Rate (HR). The legend indicates: Our model (red), Wang et al. [26] (black), and Itti et al. [18] (green).

Fig. 2. Average MM-Distance (Pixel) vs Forgetting factor. The x-axis represents the Forgetting factor (0 to 1) and the y-axis represents the Average MM-Distance (Pixel). The red dot indicates the minimum distance at a forgetting factor of 0.7.

ROC	Itti et al. [18]	Wang et al. [26]	Our model
ROC	0.13	0.13	0.13

ROC curves for different models. The x-axis represents the False Positive Rate (FPR) and the y-axis represents the Hit Rate (HR). The legend indicates: Our model (red), Wang et al. [26] (black), and Itti et al. [18] (green).

Average MM-Distance (Pixel) vs Forgetting factor. The x-axis represents the Forgetting factor (0 to 1) and the y-axis represents the Average MM-Distance (Pixel). The red dot indicates the minimum distance at a forgetting factor of 0.7.

### 3.4. Assessment of the forgetting factor

The forgetting factor  $\epsilon$  is a key parameter in the model. It is defined as the ratio of the number of hits to the total number of trials. The forgetting factor  $\epsilon$  is a key parameter in the model. It is defined as the ratio of the number of hits to the total number of trials. The forgetting factor  $\epsilon$  is a key parameter in the model. It is defined as the ratio of the number of hits to the total number of trials.

### 4. Conclusion, Discussion and Future Work

The model is evaluated using ROC curves and Average MM-Distance. The results show that the proposed model outperforms existing methods. The forgetting factor  $\epsilon = 0.7$  is identified as the optimal value for minimizing the Average MM-Distance. Future work includes extending the model to more complex scenarios and improving its robustness.

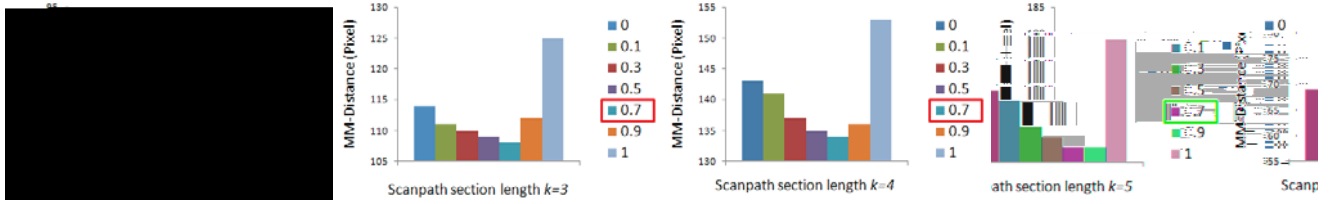


Figure 1: MM-Distance (Pixel) vs Scanpath section length k.

## 1 Introduction

1 In this paper, we propose a novel method for measuring the edit distance between two scanpaths. The edit distance is a metric that quantifies the dissimilarity between two sequences of elements. In the context of eye tracking, the scanpath is a sequence of fixations and saccades. The edit distance between two scanpaths is defined as the minimum number of insertions, deletions, and substitutions required to transform one scanpath into another. This metric is useful for comparing scanpaths across different conditions or subjects, and for identifying patterns of eye movement behavior.

## 2 Acknowledgments

1 This work was supported by the National Science Foundation (NSF) Grant IRI-1212122, and the National Eye Institute (NEI) Grant EY-15324. We thank the anonymous reviewers for their helpful comments.

## 3 References

- 1 R. A. Jacobs, S. H. Jacobs, F. E. Polzella, and S. S. Jacobs. Computer Vision and Pattern Recognition, 2001.
- 2 A. J. Jacobs, D. A. Jacobs, and S. M. Jacobs. Investigative Ophthalmology, 1991.
- 3 H. Jacobs and U. Jacobs. Neural Computation, 1991.
- 4 C. Jacobs, R. Jacobs, R. Jacobs, and R. Jacobs. Journal of Neuroscience, 2001.
- 5 N. Jacobs, S. Jacobs, and S. Jacobs. NIPS, 2001.
- 6 M. C. Jacobs, C. P. Jacobs, and M. E. Jacobs. Journal of Neuroscience, 2001.
- 7 A. C. Jacobs and V. Jacobs. Annual Review of Psychology, 1991.

## 4 Discussion

1 In this paper, we have presented a novel method for measuring the edit distance between two scanpaths. This method is based on the edit distance metric, which is a well-established metric for comparing sequences. The edit distance between two scanpaths is defined as the minimum number of insertions, deletions, and substitutions required to transform one scanpath into another. This metric is useful for comparing scanpaths across different conditions or subjects, and for identifying patterns of eye movement behavior.

## 5 Conclusion

1 In this paper, we have presented a novel method for measuring the edit distance between two scanpaths. This method is based on the edit distance metric, which is a well-established metric for comparing sequences. The edit distance between two scanpaths is defined as the minimum number of insertions, deletions, and substitutions required to transform one scanpath into another. This metric is useful for comparing scanpaths across different conditions or subjects, and for identifying patterns of eye movement behavior.

## 6 Acknowledgments

- 1 H. Jacobs, C. Jacobs, P. P. Jacobs, and S. Jacobs. NIPS, 2001.
- 2 H. Jacobs, A. Jacobs, S. Jacobs, and I. Jacobs. Proc. R. Soc. Lond. B, 1991.
- 3 H. Jacobs, S. Jacobs, A. Jacobs, and P. Jacobs. Computer Vision and Pattern Recognition, 2001.
- 4 I. Jacobs, C. Jacobs, E. N. Jacobs, and A. Jacobs. IEEE PAMI, 1991.
- 5 S. Jacobs, A. Jacobs, and I. Jacobs. Advanced in Neural Information Processing System, 1991.
- 6 I. Jacobs, P. Jacobs, and S. Jacobs. NIPS, 2001.
- 7 O. Jacobs and D. F. Jacobs. Nature, 1991.
- 8 R. Jacobs, P. Jacobs, and C. Jacobs. Journal of Vision, 2001.
- 9 S. Jacobs, M. C. Jacobs, and E. Jacobs. Journal of Statistical Physics 65: 579-16, 1991.
- 10 E. S. Jacobs, O. Jacobs, and N. Jacobs. Annual Review of Neuroscience, 2001.
- 11 F. Jacobs, P. Jacobs, and S. Jacobs. Nature Reviews Neuroscience, 2003.
- 12 Jacobs, H. Jacobs, and M. Jacobs. CVPR, 2001.