**Brain Inspired Cognitive Systems**
**August 29 – September 1, 2004**
**University of Stirling, Scotland, UK**

# EYE MOVEMENT PREDICTIONS ENHANCED BY SACCADE DETECTION

Martin Böhme, Christopher Krause, Erhardt Barth, and Thomas Martinetz

Institute for Neuro- and Bioinformatics
University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

**ABSTRACT** We present a model for predicting eye movements of an observer viewing dynamic scenes. Supervised-learning techniques are used to tune the model for a particular observer. The approach builds on earlier work [3], adding a saccade detector that is used to switch between two different algorithms for saccade and inter-saccade prediction, respectively. This separation yields a significant improvement in prediction quality. The predictor for saccade targets operates on a list of salient locations. These are obtained by evaluating the intrinsic dimension of the image using the structure tensor. Prediction of eye movements between two saccades uses a model that operates on a limited history of locations attended in the past. Both models learn by minimizing the quadratic prediction error using gradient descent. Our work is motivated by applications that involve gaze-contingent interactive displays on which information is displayed as a function of gaze direction. The approach therefore differs from standard approaches in two ways: (i) we deal with dynamic scenes, and (ii) we provide means for adapting the model to a particular observer.

## 1 INTRODUCTION

Vision is a highly active process [9, 11, 12]. Our eyes are constantly scanning the environment to centre the fovea – the highest-resolution part of the retina – over targets of interest. This sequence of eye movements is called the scan-path [11]. Its shape depends both on visual features of the scene and on search strategies. These search strategies are mostly subconscious and may vary among individuals.

For the purposes of this paper, we will distinguish between the following three types of eye movements, although there are several more [8]: (i) Saccade: the eyes move rapidly to centre the fovea over a target of interest; (ii) Fixation: eye movement is inhibited to keep the gaze on a target of interest; (iii) Smooth Pursuit: the eyes track a moving object to keep it in the same relative position on the fovea.

Work on modelling and predicting eye movements has typically been carried out on static scenes [6]; only a few authors propose models for dynamic scenes, e.g. [5]. We are interested in the latter problem because our research is motivated by applications that involve gaze-contingent displays and the guiding of eye movements [1, 2]. Also, we believe that top-down and random components have a greater influence on the scan-path for static scenes than for dynamic scenes.

A gaze predictor trained for a particular individual by using supervised-learning techniques was first described in [3]. This predictor consists of two components. The first performs a prediction based on a history of previously attended points, and the sec-

Authors' e-mail addresses:
{boehme, krause, barth, martinetz}@inb.uni-luebeck.de
Home page: http://www.inb.uni-luebeck.de

ond uses a list of salient features in the image that have the potential to attract the observer's attention. These two components are then combined linearly to obtain the predicted gaze location.

In this paper, we add a saccade detector to our model and use it to switch between the two components of the predictor instead of combining them linearly. Our motivation for this is that fixation and smooth pursuit on one hand and saccades on the other are two separate processes with quite distinct properties. The quality of our predictions should therefore improve if we model these two processes separately. Our results show that this is indeed the case.

In addition, we discuss the suitability of structure-tensor-based saliency information for saccade prediction.

The layout of this paper is as follows. Section 2 describes the saccade detector and the two components of the eye movement predictor as well as the method used to extract salient points from an image sequence. Section 3 presents the results obtained on three test video sequences. Section 4 summarizes our findings and discusses issues for future research.

# 2 METHOD

Our model for eye movements consists of three components: two predictors for saccadic and non-saccadic eye movements, which we will refer to as the S predictor and NS predictor, and a saccade detector that switches between the two predictors.

### SACCADE DETECTION

We detect that a saccade has started when the speed of the eye movement exceeds 160 degrees per second. Since the eye has already travelled a certain distance at this point, we search backwards in time to find the first point where the speed exceeded a lower threshold of 20 degrees per second and define this to be the actual starting point of the saccade. The end of a saccade is detected when the speed falls back below 20 degrees per second. Using two separate thresholds in this way makes the detector more robust than methods that use a single threshold.

### THE NS (NON-SACCADIC) PREDICTOR

The NS predictor is active in the time between two saccades and uses a history of $N$ locations attended in the past to predict the gaze point in the next time step. The predicted location $\hat{\mathrm{X}}_t = (\hat{x}_t, \hat{y}_t)$ is defined by

$$\hat{\mathrm{X}}_t = \mathrm{X}_{t-1} + \mathrm{A}_{t-1}\mathrm{P}_{t-1}.$$

$\mathrm{X}_{t-1}$ is the location in the previous time step; $\mathrm{P}_{t-1} = (\mathrm{X}_{t-2} - \mathrm{X}_{t-1}, \mathrm{X}_{t-3} - \mathrm{X}_{t-1}, \ldots, \mathrm{X}_{t-N} - \mathrm{X}_{t-1})^{\mathrm{T}}$ is the history of locations attended in the past, relative to the last known location $\mathrm{X}_{t-1}$. The $(N-1) \times 2$ matrix $\mathrm{P}_{t-1}$ is mapped by the $1 \times (N-1)$ matrix $\mathrm{A}_{t-1}$ to a displacement vector that defines the shift of the gaze point from the previous to the current time step. The matrix $A_{t-1}$ is updated continuously using supervised learning in each time step, i.e. we use an incremental learning strategy.

In the case where the last saccade ended less than $N$ time steps ago, the gaze point history contains a number of samples taken during the saccade and a number of samples taken after the saccade ended. The predictor is thus being fed with data generated by two different processes, and our experience is that this causes it to make unsatisfactory predictions.

For this reason, we apply the following modification: Let $t_{se}$ be the time step when the last saccade ended, i.e. $\mathrm{X}_{t_{se}}$ is the first sample that was classified as not belonging to the saccade. If $t_{se} > t - N$, we set $\mathrm{P}_{t-1} = (\mathrm{X}_{t-2} - \mathrm{X}_{t-1}, \ldots, \mathrm{X}_{t_{se}} - \mathrm{X}_{t-1}, \ldots, \mathrm{X}_{t_{se}} - \mathrm{X}_{t-1})$ – the samples from time steps before $t_{se}$ are replaced by $X_{t_{se}}$.

Our learning procedure is as follows: We start with $\mathrm{A} = (0, \ldots, 0)$ and apply the following update rule in each time step:

$$\mathrm{A}_t = \mathrm{A}_{t-1} + \varepsilon \mathrm{e}\mathrm{P}_{t-1}^{\mathrm{T}},$$

where $\varepsilon$ is the learning rate and $\mathrm{e} = \mathrm{X}_t - \hat{\mathrm{X}}_t$ is the prediction error. The learning rate is the distance by which the algorithm walks down the error function in the direction of the gradient $\mathrm{e}\mathrm{P}_{t-1}^{\mathrm{T}}$. We have experimented with different constant learning rates as well as with rates that were decremented exponentially. The best results, however, were obtained by estimating the optimal learning rate in each iteration and weighting this value with a constant $\alpha$. In this case, the learning rate depends on the current error and is defined by

$$\varepsilon = \alpha \frac{\mathrm{e}\mathrm{P}^{\mathrm{T}}\mathrm{P}\mathrm{e}^{\mathrm{T}}}{|\mathrm{P}^{\mathrm{T}}\mathrm{P}\mathrm{e}^{\mathrm{T}}|^2}.$$

This expression is found by a line-search method that minimizes the error on the current input.

### THE S (SACCADIC) PREDICTOR

The S predictor is used when the saccade detector detects the start of a saccade. It is fed with the gaze position $\mathrm{X}_{t_{ss}}$ at the start of the saccade and a number of salient candidate locations extracted from the $M$ most current video frames. $L$ candidate locations are extracted per frame to give a total of $M \cdot L$ locations. Their positions relative to $\mathrm{X}_{t_{ss}}$ are stored in the $(M \cdot L) \times 2$ matrix $\mathrm{C} = (\mathrm{X}_1^{\mathrm{C}} - \mathrm{X}_{t_{ss}}, \ldots, \mathrm{X}_{M \cdot L}^{\mathrm{C}} - \mathrm{X}_{t_{ss}})^{\mathrm{T}}$. The predicted location $\hat{\mathrm{X}}_{t_{se}}$ for the end of the saccade is given by

$$\hat{\mathrm{X}}_{t_{se}} = \mathrm{X}_{t_{ss}} + \mathrm{BC}.$$

B is a $1 \times (M \cdot L)$ matrix that is updated continuously using the same learning rule as the NS predictor, i.e. once we know the point $X_{t_{se}}$ where the saccade actually ended, we update B using the rule

$$\mathrm{B}_{\mathrm{new}} = \mathrm{B} + \varepsilon \mathrm{e} \mathrm{C}^{\mathrm{T}},$$

where, again, $\varepsilon$ is the learning rate and $\mathrm{e} = \mathrm{X}_{t_{se}} - \hat{\mathrm{X}}_{t_{se}}$ is the prediction error. As before, the learning rate is given by

$$\varepsilon = \beta \frac{\mathrm{e} \mathrm{C}^{\mathrm{T}} \mathrm{C} \mathrm{e}^{\mathrm{T}}}{|\mathrm{C}^{\mathrm{T}} \mathrm{C} \mathrm{e}^{\mathrm{T}}|^2},$$

where $\beta$ is a constant that weights the learning rate.

### EXTRACTION OF SALIENT FEATURES

As described in a previous paper [3], our approach to saliency is based on the concept of intrinsic dimensionality that was introduced for images in [13] and shown to be useful for modelling attention with static images in [14]. The intrinsic dimension of a signal at a particular location is the number of directions in which the signal is locally non-constant. It fulfils our requirement for an "alphabet" of image changes that classifies a constant and static region with low saliency, stationary edges and uniform regions that change in time with intermediate saliency, and popping regions that have spatial structure with high saliency. We also note that i2D regions of images and image sequences (those regions where the intrinsic dimension is at least 2) have been shown to be unique, i.e. they fully specify the image [10].

The evaluation of the intrinsic dimension is possible within a geometric approach [4] and is implemented here by using the structure tensor $\mathbf{J}$, which is well known in the computer-vision literature (see e.g. [7]).

Based on the image-intensity function $f(x, y, t)$, the structure tensor $\mathbf{J}$ is defined as

$$\mathbf{J} = \mathrm{w} * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix},$$

where subscripts indicate partial derivatives and w is a spatial smoothing kernel that is applied to the products of first-order derivatives. The intrinsic dimension of $f$ is $n$ if $n$ eigenvalues of $\mathbf{J}$ are non-zero. However, we do not need to perform the eigenvalue analysis of $\mathbf{J}$ since it is possible to derive the intrinsic dimension from the invariants of $\mathbf{J}$, which are

$$\begin{aligned} H &= \frac{1}{3}\mathrm{trace}(\mathbf{J}) && = \lambda_1 + \lambda_2 + \lambda_3 \\ S &= |\mathrm{M}_{11}| + |\mathrm{M}_{22}| + |\mathrm{M}_{33}| && = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K &= |\mathbf{J}| && = \lambda_1\lambda_2\lambda_3, \end{aligned}$$

where $\mathrm{M}_{ij}$ are the minors of $\mathbf{J}$ obtained by eliminating row $4 - i$ and column $4 - j$ of $\mathbf{J}$. The $\lambda_i$ are the eigenvalues of $\mathbf{J}$. Since $\mathbf{J}$ is a positive definite matrix, the intrinsic dimension is at least 1 if $H$ is non-zero, at least 2 if $S$ is non-zero, and 3 if $K$ is non-zero. Currently, we use only the invariant $S$ for saliency. This seems the simplest choice because $S \neq 0$ indicates an intrinsic dimension of at least 2, suppressing regions of dimension less than 2, which are redundant.

Candidate locations are obtained from $S(x, y, t)$ by applying a threshold $\theta$. For each connected region with $S$-values above the threshold, we determine the location with maximum $S$ and use it as a candidate location.

To extract salient features on different spatial and temporal scales, we construct a 4-level spatio-temporal Gaussian pyramid from the image sequence. The saliency measure $S$ is computed on each level, yielding one list of candidate locations per level. We then combine the candidates from all levels, sort them by maximum saliency and retain only the $L$ highest-saliency locations.

## 3 RESULTS

The model was trained and tested on recordings of eye movements that were made for three test video sequences. The first one is synthetic, showing a square that moves from top left to bottom right. In addition, two other squares pop in and out at different moments. The sequence runs for 30 seconds at 25 frames per second and has a resolution of 360 by

288 pixels. The second sequence is a movie trailer showing real-life scenes; it runs for 80 seconds at 30 frames per second and has a resolution of 320 by 240 pixels. The third sequence shows traffic flowing across an intersection; it runs for 15 seconds at 25 frames per second and has a resolution of 352 by 288 pixels.

The sequences were displayed on a monitor with an image size of 40 by 30 cm at a viewing distance of 50 cm, thus spanning a horizontal field of view of about 44 degrees. Eye movements were recorded at 240 samples per second using the commercial videographic eye tracker iViewX produced by SensoMotoric Instruments GmbH.

For each test sequence, recordings were made for a number of test subjects: four subjects (one recording each) for the first sequence; five subjects (one recording each) for the second sequence; and six subjects (three recordings each) for the third sequence. Unless stated otherwise, we show averages of the results for all subjects, thus reducing the influence of outliers and variations among individuals.

## NS PREDICTOR

As a baseline for evaluating the NS predictor, we used a simple model defined by

$$\hat{\mathrm{X}}_t = \mathrm{X}_{t-1},$$

i.e. the predicted gaze location for the current time step is just the actual gaze location in the previous time step. We also compared the NS model to the M2 model described in [3]. M2 is mostly identical to NS except that it does not prevent the mixing of saccadic and non-saccadic data in its history buffer since it does not contain a saccade detector.

In all tests, the constant $\alpha$, which scales the learning rate, was set to $\alpha = 0.001$. For the size of the history buffer $N$, we tested a range of values between 2 and 100; at 240 samples per second, this corresponds to a range of approximately 8 to 400 milliseconds.

Figure 1 shows the average squared prediction error for the three models. On the first sequence, we note that NS performs better than M1 for all history lengths, with decreasing error for increasing history length. On the second sequence, NS and M1 show approximately the same error, independent of the history length. This is because the first sequence induces smooth pursuit movements in the test subjects, whereas the subjects typically performed only saccades and fixations on the second sequence. On the third sequence, where the traffic flow provoked smooth pursuit movements, we again see an improvement relative to M1 that increases with history length.

Comparing NS and M2, we find that for small $N$, they show roughly similar results, but for larger $N$, the performance of M2 starts to degrade quite rapidly. This shows that avoiding the mixing of saccadic and non-saccadic data in the history buffer is critical for achieving accuracy and robustness, providing a strong argument for the separation of the two predictors. The strength of this effect increases with the size of the history buffer.

Figure 4 shows the cumulated error over time relative to M1 for a single subject (M. B.) on the first sequence. The plot shows certain time intervals where the NS error decreases steadily relative to M1, and others where it remains constant. The phases that show a steady decrease correspond to smooth pursuit movements made by the test subject. This shows that NS has an advantage over M1 primarily for smooth pursuit movements, whereas both models perform similarly on fixations. This is to be expected because the prediction that M1 makes is just an idealized description of a fixation; the small errors that it makes are due to small random movements that the eyes make even during a fixation.

## S PREDICTOR

To evaluate the S predictor, we computed the ratio between the average squared prediction error and the average squared saccade length. A ratio of less than 1 means that, on average, the predictions moved in the right direction relative to the starting point of the saccade.

As an aid for assessing the results, we also computed the errors made by three other predictors. The first simply chooses a random point in the image as the predicted saccade target. Of course, we do not expect this predictor to perform well, but we include it as a baseline reference. The second predictor chooses the location with the maximum saliency among the $M \cdot L$ candidate locations. The third, a hypothetical "ideal" predictor, always picks the candidate location that lies closest to the actual end point of the saccade. The benefit of this "ideal" predictor is twofold: First, it gives us a bound for the best prediction result we can expect on the available information if we assume a predictor that selects one of the candidate locations, without any averaging between locations. Second, it gives us a tool for evaluating the quality of the candidate locations generated by our saliency measure.

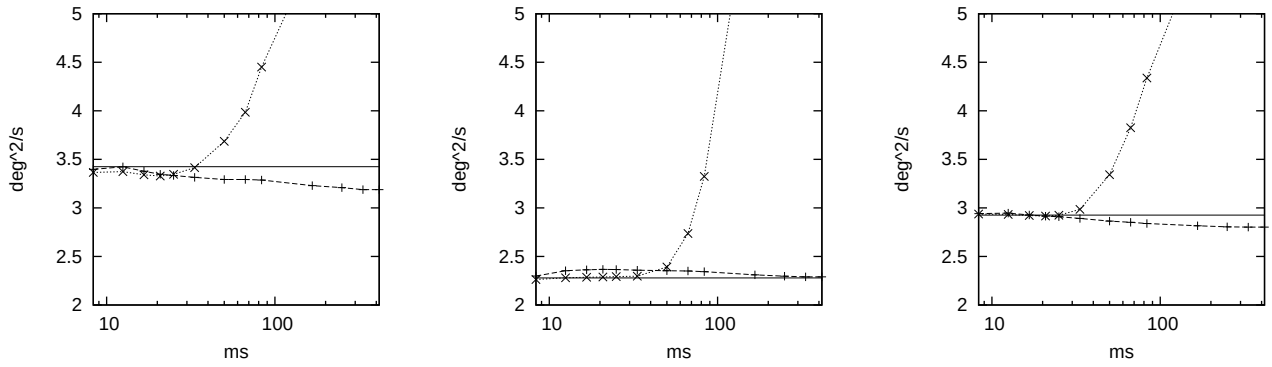


**Figure 1:** Squared prediction error for the M1 (solid), M2 (dotted) and NS (dashed) predictors on the three test sequences (synthetic sequence, movie trailer and traffic scene, from left to right). The horizontal axis plots the history length in milliseconds.

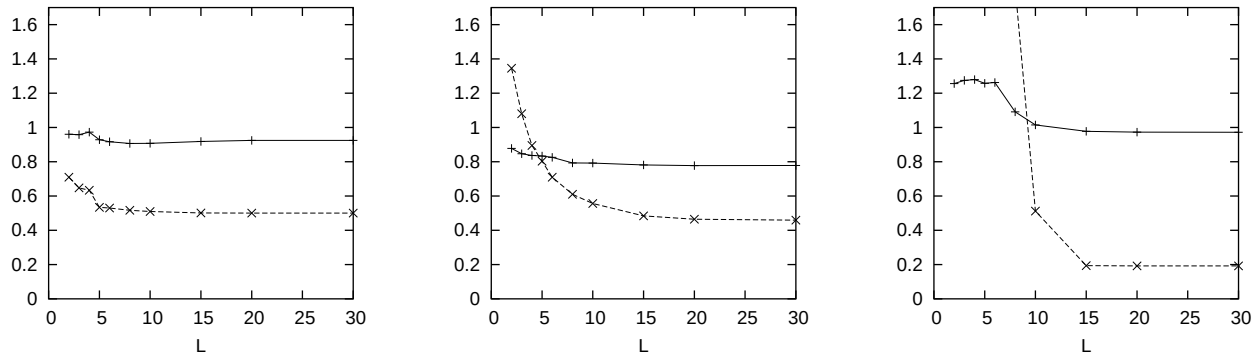


**Figure 2:** Ratio of the average squared prediction error and average squared saccade length for the three sequences (synthetic sequence, movie trailer and traffic scene, from left to right), for the S predictor (solid) and the "ideal" predictor (dashed). The horizontal axis plots $L$, the number of salient features per frame; $M$, the number of video frames from which features were extracted, was set to 1.

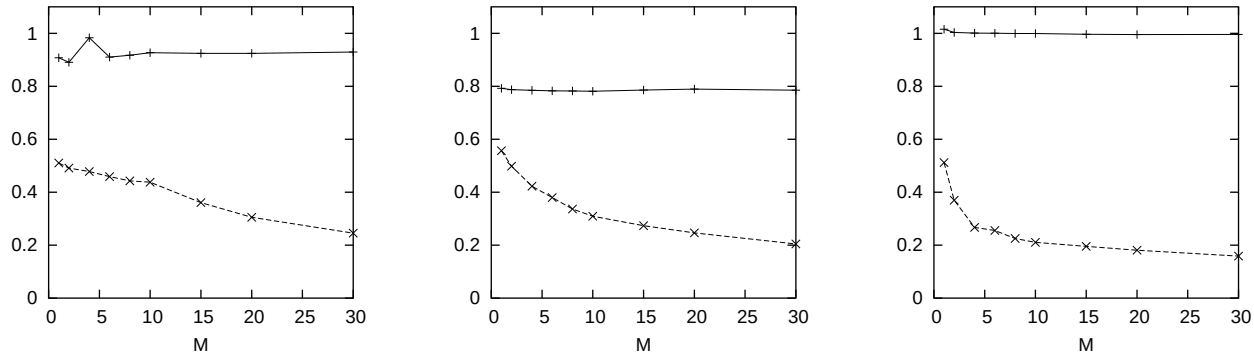


**Figure 3:** Relative prediction error as in Figure 2 for the S predictor (solid) and the "ideal" predictor (dashed). The horizontal axis plots $M$, the number of video frames; $L = 10$ salient features were extracted from each frame.

For all results presented here, the constant $\beta$, which scales the learning rate, was set to $\beta = 0.05$; we found that this value gave the best results. The threshold $\theta$ for the saliency was set to 0.5 of the maximum in the current frame.

Figure 2 shows the results for the S predictor (solid) and the "ideal" predictor (dashed) for different numbers of features $L$ extracted from a single video frame. For all three sequences, the S predictor achieves a relative error of less than 1 for large $L$, showing that, on the average, the predictions moved in the right direction relative to the saccade starting point. The relative errors of 0.2 to 0.5 achieved by the "ideal" predictor show the potential for improvement that exists with the given saliency information.

The results for the "random" and "maximum saliency" predictors are not shown in the plots since they are constant for all $L$. The "random" predictor yields relative errors of 2.1, 2.9 and 3.6 on the three sequences. The "maximum saliency" predictor yields errors of 0.94, 2.0 and 5.0. This shows that always choosing the most salient location is not a good general strategy – it yields results comparable to the S predictor on the synthetic sequence but performs badly on the two natural sequences. We believe this is because the synthetic sequence produces only a few salient candidate locations per frame, so that the most salient location has a good chance of being the right choice. The natural sequences, on the other hand, produce more candidate locations, and apparently there are more criteria than magnitude of saliency alone that drive the saccade target selection process.

Figure 3 again shows the results of the S predictor and the "ideal" predictor, this time for a varying number of video frames $M$, with a constant number of features $L = 10$ extracted from each frame. While the "ideal" predictor shows the expected decrease in the error with increasing $M$, the error for the S predictor remains relatively constant. We conclude that the S predictor does not learn the relative significance of salient features with different "ages" effectively.

# 4 CONCLUSIONS AND OUTLOOK

Using a saccade detector to switch between separate predictors for saccadic and non-saccadic eye movements improves prediction results significantly. The rationale is that saccadic and non-saccadic eye movements are two distinct processes that should be modelled separately.

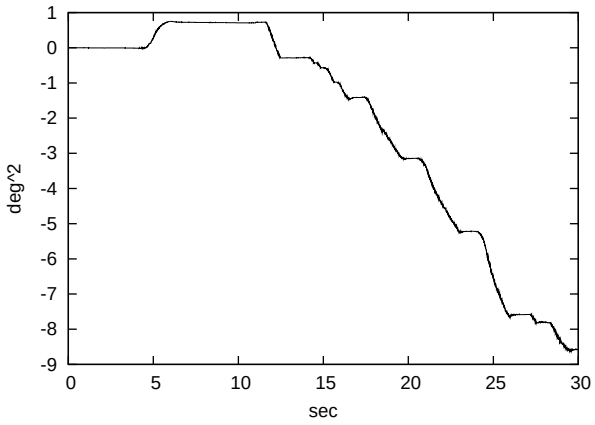


**Figure 4:** Cumulated NS error over time relative to M1 for a single subject on the first sequence. The horizontal axis plots elapsed time in seconds.

Our current implementation of the saccade predictor allows several candidate locations to be mixed or averaged together to give the predicted location. This is reasonable if there are several candidate locations that lie close together; but in other cases, where the candidate locations lie far apart, the mixing effect is not desirable. Indeed, the results of the predictor are not yet satisfactory – the predictor moves in the right direction the majority of the time but often underestimates saccade length due to the mixing effect. With an average error of about 0.9 of the average saccade length, we have covered only ten percent of the way to a perfect predictor.

Of course, the processes that determine saccade targets in the human optical system probably do not involve a linear mixing of candidate locations but rather a selection of one of these locations. For this reason, we see the current model only as a first step and intend to experiment with other predictors that select one of the candidate points as the most likely saccade target, based on criteria such as the distance to the candidate locations and the magnitude of the $S$ measure. We have demonstrated the potential of this kind of predictor using an "ideal" predictor that always picks the candidate location that is closest to the actual saccade target. The good results obtained using this "best case" analysis also demonstrate the suitability of our saliency measure for generating candidate locations (note that the $L \leq 30$ candidate locations selected in each frame represent less than 0.04% of the possible pixel locations).

The saccade predictor that we present in this paper predicts a saccade target given that we know a saccade is taking place. Ultimately, though, we want to be able to predict *that* a saccade will take place before it starts. To do this, a predictor will need to be able to respond to temporal transients in the saliency measure, i.e. it needs to know the absolute magnitude of the saliency measure for the candidate locations. The predictor we have used so far only knows the relative saliency of the candidates (implicitly, through their position in the candidate vector).

Finally, we are interested in training a model to predict higher-level strategies and phenomena such as inhibition-of-return (a bias that tends to inhibit saccades to recently attended locations). To do this, we suggest providing the model with a list of the last few fixations and the durations for which they were held.

We conclude that the nonlinear decoupling of pursuit and saccade predictors allows for a good prediction during pursuit based on a linear model. For saccade prediction, however, the nonlinear saliency measure combined with a linear combination rule remains unsatisfactory and may be improved by using a nonlinear selection rule.

## ACKNOWLEDGEMENTS

*Information technology for active perception (Itap)* [1] research is supported by the German Ministry of Education and Research (BMBF) as part of the interdisciplinary project *ModKog* (grant number 01IBC01B). We thank SensoMotoric Instruments GmbH for their eye-tracking support; data were obtained using their iViewX system.

## REFERENCES

[1] Information technology for active perception (Itap). `http://www.inb.uni-luebeck.de/Itap`.

[2] E. Barth, J. Drewes, and T. Martinetz. Dynamic predictions of tracked gaze. In *Seventh International Symposium on Signal Processing and its Applications, Special Session on Foveated Vision in Image and Video Processing*, 2003.

[3] E. Barth, J. Drewes, and T. Martinetz. Individual predictions of eye-movements with dynamic scenes. In B. Rogowitz and T. Pappas, editors, *Electronic Imaging 2003*, volume 5007. SPIE, 2003.

[4] E. Barth and A. B. Watson. A geometric framework for nonlinear visual coding. *Optics Express*, 7:155–165, 2000. `http://www.opticsexpress.org/oearchive/source/23045.htm`.

[5] G. Boccignone, A. Marcelli, and G. Somma. Analysis of dynamic scenes based on visual attention. In *Proceedings of AIIA 2002*, Siena, Italy, 2002.

[6] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[7] B. Jähne, H. Haußecker, and P. Geißler, editors. *Handbook of Computer Vision and Applications*. Academic Press, Boston, 1999.

[8] R. J. Leigh and D. S. Zee. *The Neurology of Eye Movements*. Oxford University Press, New York, third edition, 1999.

[9] D. M. MacKay. *Behind the Eye*. Basil Blackwell, Oxford, 1991.

[10] C. Mota and E. Barth. On the uniqueness of curvature features. In G. Baratoff and H. Neumann, editors, *Dynamische Perzeption*, volume 9 of *Proceedings in Artificial Intelligence*, pages 175–178, Cologne, 2000. Infix Verlag.

[11] D. Noton and L. Stark. Eye movements and visual perception. *Scientific American*, 224(6):34–43, 1971.

[12] J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–1031, 2001.

[13] C. Zetzsche and E. Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. *Vision Research*, 30:1111–1117, 1990.

[14] C. Zetzsche, K. Schill, H. Deubel, G. Krieger, E. Umkehrer, and S. Beinlich. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In R. Pfeifer, B. Blumberg, J.-A. Meyer, and S. W. Wilson, editors, *From Animals to Animates: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, volume 5, pages 120–126. MIT Press, Cambridge, 1998.