Eye Movement Predictions on Natural Videos

Martin Böhme [*], Michael Dorr, Christopher Krause,
Thomas Martinetz and Erhardt Barth

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

Abstract

We analyze the predictability of eye movements of observers viewing dynamic scenes. We first assess the effectiveness of model-based prediction. The model is divided into inter-saccade prediction, which is based on a limited history of attended locations, and saccade prediction, which is based on a list of salient locations. The quality of the predictions and of the underlying saliency maps is tested on a large set of eye movement data recorded on high-resolution real-world video sequences. In addition, frequently fixated locations are used to predict individual eye movements to obtain a reference for model-based predictions.

Key words:
Eye movement prediction; saccade prediction; scan-path; saliency maps

1 Introduction

Vision is a highly active process [1–3]. 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 [2]. Its shape depends both on visual features of the scene and on scanning strategies. These scanning strategies are mostly subconscious and may vary among individuals.

[*] Corresponding author.
Email addresses: boehme@inb.uni-luebeck.de (Martin Böhme), dorr@inb.uni-luebeck.de (Michael Dorr), krause@inb.uni-luebeck.de (Christopher Krause), martinetz@inb.uni-luebeck.de (Thomas Martinetz), barth@inb.uni-luebeck.de (Erhardt Barth).

Article published in Neurocomputing ***69****, Issues 16–18, 2006, pp. 1996–2004*
http://www.elsevier.com/wps/product/cws_home/505628

For the purposes of this paper, we will distinguish between the following three types of eye movements, although there are several more [4]: (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 [5,6]; only recently has there been increased interest in models for dynamic scenes, e.g. [7–9]. We are interested in the latter problem because our research is motivated by applications that involve gaze-contingent displays and the guidance of eye movements [8,10]. Also, we believe that top-down and random components have a greater influence on the scan-path for static scenes, whereas the influence of bottom-up factors is greater for dynamic scenes. We cite several arguments in support of this: First, motion and temporal change (which are, of course, only present in dynamic scenes) have been shown to be stronger predictors of human saccades than colour, intensity or orientation [9]. Second, smooth pursuit, an involuntary eye movement controlled by bottom-up mechanisms, is only observed on dynamic scenes. Third, the human visual system evolved in an environment where fast reflexive reactions to dynamic visual cues were important.

Our approach to gaze prediction divides the problem into two parts: First, predicting the eye movements made between saccades; and second, predicting the targets of saccades. We will refer to these two parts of the problem as intersaccadic prediction and saccade prediction.

Of course, dividing the problem in this way requires some means of switching between intersaccadic prediction and saccade prediction. Currently, we use a saccade detector, i.e. we switch from intersaccadic prediction to saccade prediction once we detect that a saccade is taking place. Ultimately, one would also want to predict *that* a saccade will take place before it actually starts.

For intersaccadic prediction, we use a predictor based on supervised-learning techniques that uses a history of previously attended locations to predict the gaze position in the next time step.

Saccade prediction is certainly the harder of the two subproblems. Like other authors [5,7,11–13], we base our approach on a *saliency map* that assigns a certain degree of saliency to every location in every frame of a video sequence. Various techniques exist for computing saliency maps, but they are all based, in one way or another, on local low-level image properties such as contrast, motion or edge density, and are intended to model the processes in the human visual system that generate saccade targets.

From a saliency map and a history of previously attended locations, one would ideally want to be able to predict a single location that has a high probability of being the next saccade target. However, we fear that this goal may be unattainable. We believe that the human visual system uses low-level features such as those used in saliency maps to generate a list of candidate locations for the next saccade target, and that top-down attentional mechanisms then select one of the candidate locations as the actual saccade target. This selection mechanism is probably very difficult to model algorithmically.

In our view, a more realistic goal is therefore to predict a certain number of candidate locations – say, five to ten – that will with high probability include the actual saccade target, and our results show that a small number of target locations usually covers most of the variations in the eye movements made by different observers.

In spite of our reservations, we have also attempted to see how well we can predict a single "most likely" saccade target based on a list of candidate locations, and we will report on these findings also. However, for our purposes [10], the prediction of a few candidate locations suffices.

The layout of this paper is as follows. Section 2 describes our methods for intersaccadic prediction, computation of saliency maps, extraction of candidate locations for saccade targets, and the prediction of a single "most likely" saccade target. Section 3 compares the results of our methods with the eye movements made by test subjects on 18 high-resolution, real-world video sequences. Section 4 summarizes our findings and discusses issues for future research.

2 Method

Since our approach splits up the prediction of eye movements into intersaccadic prediction and saccade prediction, we first describe the saccade detector that is used to switch between these two modalities. We then present the method used for intersaccadic prediction. Next, we introduce a saliency measure based on the concept of intrinsic dimensionality as well as an "empirical" saliency measure computed from the eye movements actually made by the test subjects; the empirical saliency measure will be used as a baseline for assessing the quality of the analytical saliency measure. We then present an algorithm for extracting individual candidate locations for saccade targets from a saliency map. Finally, we describe a predictor that predicts a single saccade target from a list of such candidate locations.

2.1 Saccade Detection

Saccades are detected in a two-step procedure. To initialize the search for a saccade onset, gaze velocity has to exceed a high threshold $\theta_1 = 150°/\text{s}$ first. Then, going back in time, saccade onset is defined as the point in time where the velocity exceeds a lower threshold $\theta_0 = 20°/\text{s}$ that is biologically more plausible but less robust to noise. Saccade offset is reached when gaze velocity falls below θ_0 again. To further improve robustness against impulse noise, we check the resulting saccades for a minimum and maximum length (15 ms and 120 ms, respectively) and a peak velocity of less than 1000°/s.

The thresholds used in the saccade detection algorithm were based on biologically plausible values [14]; these were then fine-tuned to optimize the algorithm's performance, which was validated by comparing against hand-labelled saccades on the same data. A few saccades of low amplitude were not detected by the saccade detector, but in some of these cases, it was difficult to determine even by visual inspection whether the data really constituted a saccade or just eye tracker noise. In any event, we are not concerned that the omission of a few low-amplitude saccades would significantly affect our results since such low-amplitude saccades typically do not shift the gaze away from the previously fixated object.

2.2 Intersaccadic Prediction

The intersaccadic 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

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

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 X_{t-1}. The $(N-1) \times 2$ matrix P_{t-1} is mapped by the $1 \times (N-1)$ matrix 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 ε 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 α. Thus, we find the ε that minimizes

$$\|\mathrm{X}_t - (\mathrm{X}_{t-1} + \mathrm{A}(\varepsilon)\mathrm{P}_{t-1})\|_2^2,$$

where $\mathrm{A}(\varepsilon) = \mathrm{A}_{t-1} + \varepsilon \mathrm{e}\mathrm{P}_{t-1}^{\mathrm{T}}$, and weight this value with α, obtaining

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

We note that prior knowledge about the statistics of intersaccadic eye movements could have been used in the design of the intersaccadic predictor. A Kalman filter would have been an obvious choice, and Kalman filtering has indeed been used to implement smooth pursuit in active vision systems [15]. However, we are interested in the quality of the results that can be achieved by a supervised-learning algorithm that makes few assumptions about the underlying process.

2.3 Saliency Map Generation

As described previously [16,17], our approach to saliency is based on the concept of intrinsic dimensionality that was introduced for images in [18] and was shown to be useful for modelling attention with static images in [19]. 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 transient patterns that have spatial structure with high saliency. We also note that those regions of images and image sequences where the intrinsic dimension is at least 2 have been shown to be unique, i.e. they fully specify the image [20].

The evaluation of the intrinsic dimension is possible within a geometric approach that is plausible for biological vision [21] and is implemented here by using the structure tensor **J**, which is well known in the computer-vision literature (see e.g. [22]).

Based on the image-intensity function $f(x, y, t)$, the structure tensor **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 **J** are non-zero. However, we do not need to perform the eigenvalue analysis of **J** since it is possible to derive the intrinsic dimension from the invariants of **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 M_{ij} are the minors of **J** obtained by eliminating row $4 - i$ and column $4 - j$ of **J**. The λ_i are the eigenvalues of **J**. Since **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.

To extract salient features on different spatial and temporal scales, we construct a 4-level spatio-temporal Gaussian pyramid from the image sequence and compute the saliency measures on each level. Such a pyramid is constructed from the image sequence by successive blurring and sub-sampling.

2.4 Empirical Saliency Maps

As a baseline for assessing the saliency maps computed analytically, as described above, we use *empirical* saliency maps – i.e. saliency maps computed from the actual eye movements of the test subjects. In a sense, they give us an idea of what the "ideal" saliency map should look like for a given video sequence, and they can serve as a basis for judging what the best possible results are that we can expect for predictions made solely on the basis of a saliency map generated from the image data, without taking individual top-down strategies into account.

To generate the empirical saliency map for a video frame, we determine the current gaze position of each observer and place a Gaussian with a standard deviation of $\sigma = 16$ pixels at each of these positions. The superposition of these Gaussians then yields the empirical saliency map.

2.5 Salient Locations

In previous work [16,23], we extracted salient locations from the saliency map by applying a threshold of 0.5 of the maximum saliency in the map. For each connected region with saliency values above the threshold, we extracted one salient location by determining the location with maximum saliency. The robustness of this approach proved to be unsatisfactory. For example, if a small region in the saliency map has values substantially higher than the rest of the map, all of the map except for the high-saliency region will be suppressed.

For this reason, we developed a new, more robust approach [17], based on the mechanism of "selective lateral inhibition". The idea is to avoid two salient locations being generated closer together than a certain distance. Therefore, when a location has been extracted, we attenuate the saliency values of points around the location using an inverted Gaussian to inhibit the generation of further salient locations close to the existing location.

The following algorithm uses the mechanism of selective lateral inhibition to extract n salient locations from the image in the order of decreasing saliency:

```
$\mathcal{S}_1 = \mathcal{S}$
for $i = 1 \dots n$ do
    $(x_i, y_i) := \operatorname*{argmax}_{(x,y)} \mathcal{S}_i(x, y)$
    $\mathcal{S}_{i+1}(x,y) := \begin{cases} \mathcal{S}_i(x,y) \cdot G(x - x_i, y - y_i) & x_i - W < x < x_i + W \text{ and} \\ & y_i - W < y < y_i + W \\ \mathcal{S}_i(x,y) & \text{otherwise} \end{cases}$
end for
```

where $\mathcal{S}$ is the saliency measure (one of H, S or K), $G(x, y) = 1 - \mathrm{e}^{-\frac{x^2+y^2}{2\sigma^2}}$ is an inverted radial Gaussian of width σ, and $W = 2\sigma$ is the window width. We thus repeatedly find the point with maximum saliency and then attenuate the saliency values around this point. The extracted locations are then just the (x_i, y_i).

Note that the Gaussian is truncated with a square window, even though a circular window would have agreed better with the radial shape of the Gaussian. We believe this is an acceptable optimization because the additional values that are included in the square window beyond those that a circular window would contain are close to 1 and the exact shape of the window does not have a significant impact on the behaviour of the algorithm.

When extracting salient locations on different levels of the spatio-temporal pyramid, the width of the Gaussian in pixels is kept constant for all levels. Effectively, this increases the width of the Gaussian relative to the image the more the resolution of the image is decreased. The idea is that the lower-resolution levels of the image capture coarser structure, which should thus be suppressed using a kernel of greater width.

2.6 Saccade Prediction

From a list of salient locations computed using the algorithm described above, the saccade predictor attempts to predict a single most probable saccade target. 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 recent 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 $\mathrm{C} = (\mathrm{X}^{\mathrm{C}}_1 - \mathrm{X}_{t_{ss}}, \ldots, \mathrm{X}^{\mathrm{C}}_{M \cdot L} - \mathrm{X}_{t_{ss}})^{\mathrm{T}}$. The predicted location $\hat{\mathrm{X}}_{t_{se}}$ for the end of the saccade is given by

$$\hat{\mathrm{X}}_{t_{se}} = \mathrm{X}_{t_{ss}} + \mathrm{B}\,\mathrm{C}.$$

B is a $1 \times (M \cdot L)$ matrix that is updated continuously using the same learning rule as the intersaccadic 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, ε is the learning rate and $\mathrm{e} = \mathrm{X}_{t_{se}} - \hat{\mathrm{X}}_{t_{se}}$ is the prediction error. As for the intersaccadic predictor, 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 β is a constant that weights the learning rate.

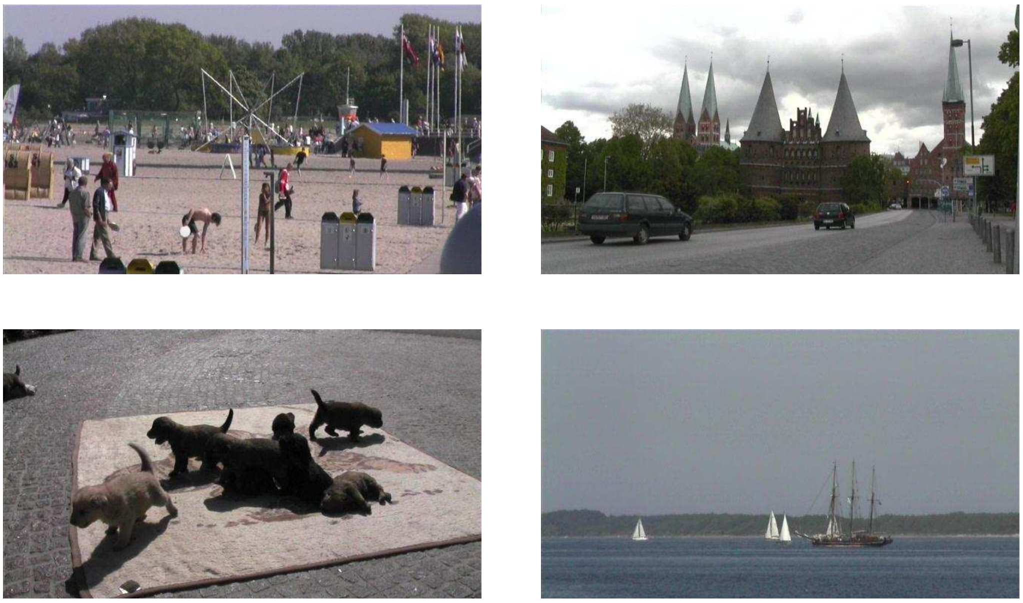

Fig. 1. Still frames from selected video sequences

3 Results

The methods described in Section 2 were tested on recordings of eye movements that were made for 18 test video sequences. The video sequences were recorded using a JVC JY-HD10 HDTV video camera and depicted a variety of real-world scenes in and around Lübeck. (7 sequences: people in a pedestrian area, on the beach, playing in a park; 4 sequences: populated streets and roundabouts; 4 sequences: animals; 3 sequences: scenes of almost still life character, e.g. a ship passing by in the distance. Still frames from some of the sequences are shown in Figure 1.) Each video sequence had a length of 20 seconds, a resolution of 1280 by 720 pixels (aspect ratio 16:9) and a frame rate of 30 frames per second (progressive scan). In general, the video camera was fixed for these recordings; only a few sequences contained small amounts of camera movement. If the sequences had contained strong camera movement, a video stabilization algorithm could have been employed.

The sequences were displayed on a monitor with an area of 40 by 30.6 cm, the video occupied an area of 39.8 by 22.8 cm, and the parts of the screen not occupied by video were black. The viewing distance was 45 cm, the video thus spanned a horizontal field of view of about 48 degrees. The test subjects were instructed to watch the sequences attentively; no other specific task was given. Eye movements were recorded at 250 samples per second using the commercial videographic eye tracker Eyelink II produced by SR Research. For each test sequence, recordings were made for 54 test subjects. The eye tracker flags invalid samples (the usual reason for these is that the subject blinks). Recordings that contained more than 5% invalid samples were discarded, leaving between 37 and 52 recordings per video sequence.

3.1 Intersaccadic Prediction

As a baseline for evaluating the intersaccadic (IS) predictor, we used a simple model, which we will refer to as M1, 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. In many cases – namely during fixations – this is in fact already an almost ideal model of intersaccadic eye movements. We also compared the IS model to the M2 model described in [16]. M2 is mostly identical to IS except that it does not prevent the mixing of saccadic and non-saccadic data in its history buffer.

The constant α, 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 1 and 100; at 250 samples per second, this corresponds to a range of 4 to 400 milliseconds. Because the video sequences are relatively short, the A matrix in the intersaccadic predictor was not reset between different test subjects. This provided the predictor with more training samples but reduced the potential benefit of tuning the predictor for an individual observer.

Figure 2 shows the mean squared prediction error for the three models, averaged over all sequences and all test subjects; confidence intervals (confidence value 99.9%) are also shown. The IS model performs consistently better than the baseline M1 model for all history lengths, though the error made by IS increases for large history lengths (beyond around 100 ms). This is in contrast to our previous results reported in [23], where the prediction error decreased continuously with increasing history length. We speculate that this may be because the prediction of smooth pursuit movements in particular seems to benefit from large history lengths, and our current, more natural, set of test sequences does not induce nearly as much smooth pursuit as the test sequences used in [23]. As the error bars show, the differences between the IS and M1 predictors are statistically significant for all history sizes of 8 ms and greater.

The M2 model shows results comparable to IS for small history sizes, but the prediction error starts to grow rapidly as the history length increases beyond about 30 ms. Since the M2 model is similar to IS except that it does not avoid the mixing of saccadic and non-saccadic data in the history buffer, we conclude that this feature is critical for achieving accuracy and robustness.

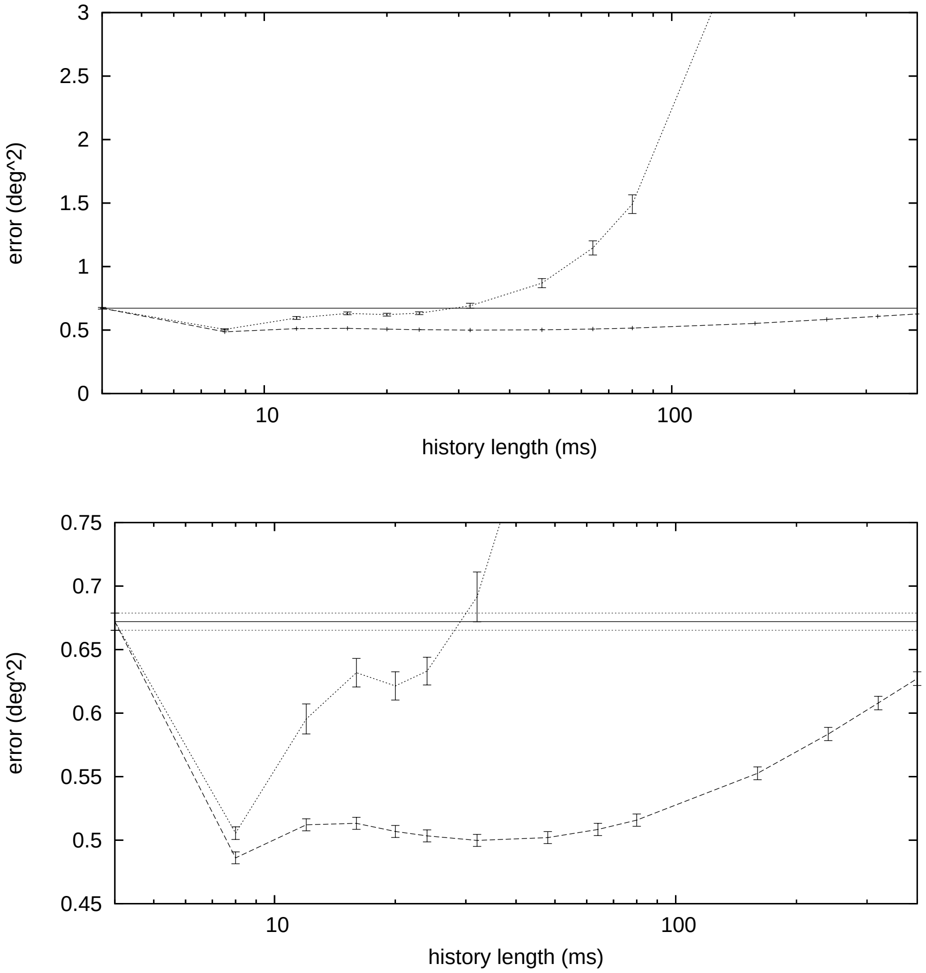

Fig. 2. Squared prediction error for the M1 (solid), M2 (dotted) and IS (dashed) predictors, averaged over all video sequences and test subjects. The horizontal axis plots the history length in milliseconds. Both graphs show the same data, but the lower graph has been stretched vertically to show more detail. The error bars show 99.9% confidence intervals for the mean error; for legibility, the top graph plots error bars only for the M2 predictor. Dotted lines indicate the confidence interval for the M1 predictor because the errors made by this predictor are independent of history length.

3.2 Salient Locations

We turn now to evaluating the quality of the candidate locations generated from the saliency maps. For every saccade made by a test subject, we extracted L candidate locations from the saliency map for the video frame that was being displayed when the saccade started. We then computed the distance from the saccade target to the closest of these candidate locations. In addition to the H, S and K saliency measures, we also used the empirical saliency measure (to quantify the best results one could expect using L candidate points) as well as L points on the image chosen completely at random (to quantify the result one would get by "just guessing"). The width of the Gaussian in the salient-location extraction algorithm was set to $\sigma = 30$ pixels.

The plot at the top of Figure 3 shows the error histogram obtained for $L = 1$, i.e. a single candidate location. The analytical saliency measures - H, S, K - show roughly comparable error, and all three perform significantly better than locations chosen at random. This is especially apparent on the left-hand side of the graph, where the H, S and K measures peak earlier than the random locations, i.e. they produce a greater number of small errors. They also produce a smaller number of large errors, but here the difference is much less marked. The empirical saliency measure shows a strong peak at about 2 degrees of error, but here, too, the graph falls off only rather slowly towards large error magnitudes. We conclude that a single candidate location does not sufficiently capture the variation among individuals' saccade targets.

It could be objected that the empirical saliency map generates unrealistically good results because it contains eye movement data from the very observers it is being tested against, and that a more realistic test might use a "leave-one-out" strategy (i.e. test every observer against an empirical saliency map constructed from the remaining observers' eye movements). This is a valid point. However, our intention was to use the same saliency map for all observers (as for the analytical saliency maps), and we note that salient candidate locations tend to be generated at points attended simultaneously by several observers; leaving one of these observers out would, in general, not make a great difference. At any rate, since the empirical saliency is only intended to give us an estimate of the best result we can expect from a saliency map, we can permit it to be slightly optimistic.

The bottom plot of Figure 3 shows the error histogram for $L = 10$ candidate locations. Again, the H, S and K saliency measures perform quite similarly and peak much earlier than the graph for random locations. The graphs for H, S and K also fall off more rapidly than the graph for the random locations up to an error of about 12 degrees, but the random locations produce fewer errors beyond this point. The reason for this is that the H, S and K candidate locations may concentrate in certain areas of the image, leaving others free. If an observer's gaze does fall into one of these free areas, a very large error results. The random candidate locations, on the other hand, tend to be distributed uniformly across the image, making exceedingly large errors unlikely.

For the empirical saliency measure, we again note a strong peak at about 2 degrees of error, but unlike for $L = 1$, the graph also falls off quickly, with only relatively few errors above about 5 degrees. We conclude that, in principle, ten candidate locations are sufficient for capturing most of the variation among individuals' saccade targets. (For lower resolution and smaller field of view, fewer points would suffice.) The analytical saliency measures H, S and K are

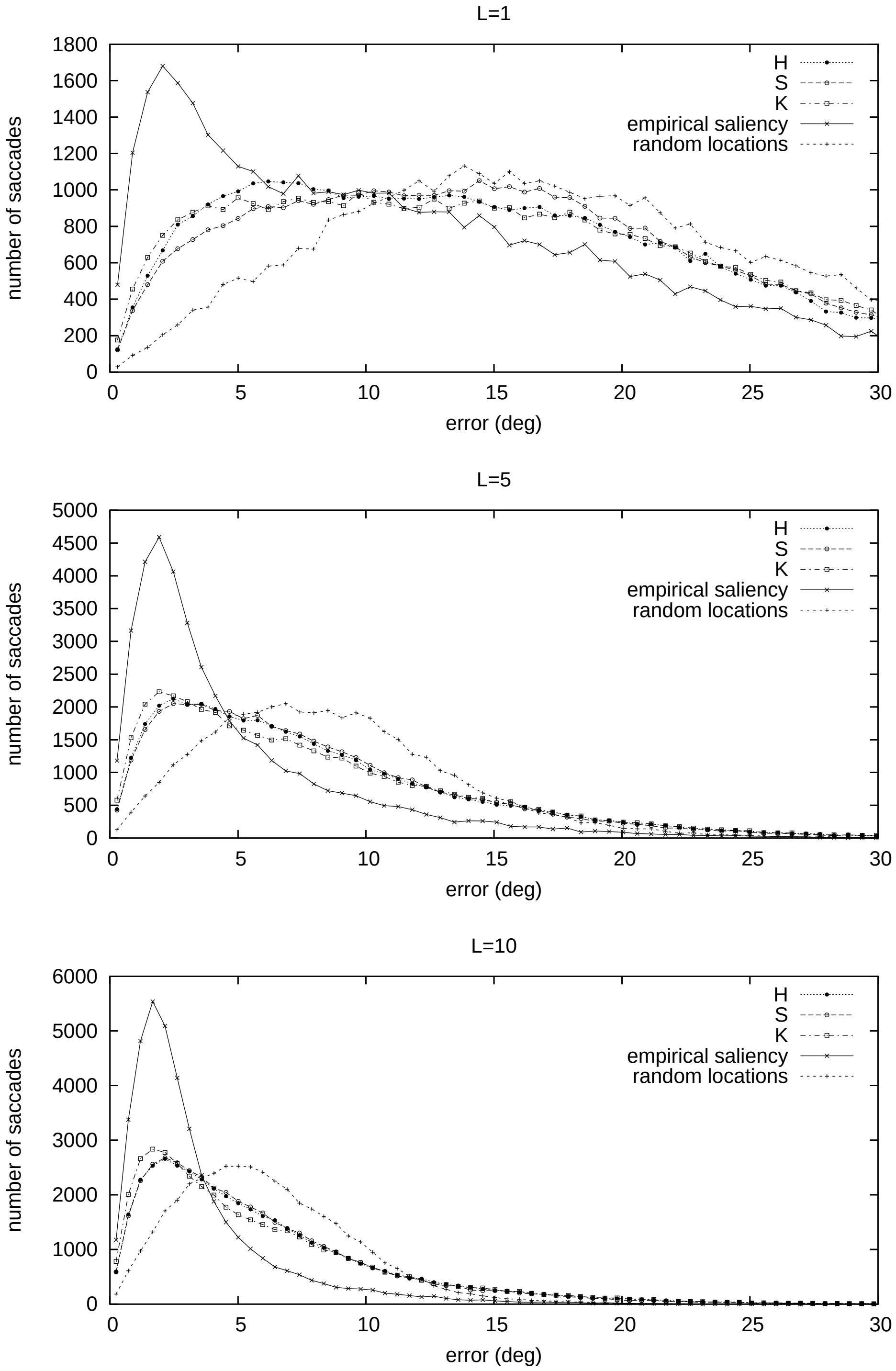

Fig. 3. Histogram of error (distance of saccade target to closest salient candidate location) for $L = 1$, $L = 5$ and $L = 10$ candidate locations. The horizontal axis plots the error magnitude in degrees, the vertical axis plots the number of saccades per histogram bin. Plots are shown for the analytical H, S and K saliency measures, the empirical saliency measure, and locations chosen at random. Results for the H, S and K measures were averaged over all levels of the spatio-temporal pyramid.

still quite far away from the optimum represented by the empirical saliency measure but nevertheless predict saccade targets well at least some of the time. Picking a winner among the three measures is difficult. The results appear to suggest that K may perform slightly better than H or S, but the differences are too small to make a definite assessment. For simpler videos with lower resolution we had found that the higher the intrinsic dimension, the better the prediction [17].

Finally, we turn to the middle plot of Figure 3, which shows the error histogram for $L = 5$ candidate locations. It presents an intermediate situation between the two plots discussed so far, but we note that the curves are more similar in shape to those for $L = 10$ than those for $L = 1$, though they do not fall off quite as quickly. This suggests that even five locations may already provide adequate coverage of the eye movements made by different individuals.

3.3 Saccade Prediction

Finally, we turn to the results of the saccade predictor, which predicts a single saccade target from a list of salient candidate locations. As an aid for assessing the results, we compare this predictor with the following three others. The first simply chooses a random point in the image as the predicted saccade target (this corresponds to the single random candidate location in the previous section). The second predictor chooses the location with maximum saliency among the candidate locations. The third, a hypothetical "ideal" predictor, always picks the candidate location that is closest to the actual saccade target. This predictor thus gives us a bound for the best prediction result we can expect on the given candidate locations if we assume a predictor that selects one of the locations, without any averaging between locations.

The constant β in the saccade predictor, which scales the learning rate, was set to $\beta = 0.05$. The parameter M, which controls the number of video frames used for prediction, was set to $M = 1$, as incorporating information from previous video frames did not seem to improve the results much. Again, because the video sequences are relatively short, the B matrix in the predictor was not reset between different test subjects. Salient candidate locations were generated using the K saliency measure evaluated on the second level of the spatio-temporal pyramid, which performed slightly better than the other levels.

Figure 4 shows the performance of the predictor for varying numbers of candidate locations L, plotted on the horizontal axis. The vertical axis plots the average squared prediction error relative to 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. Results were

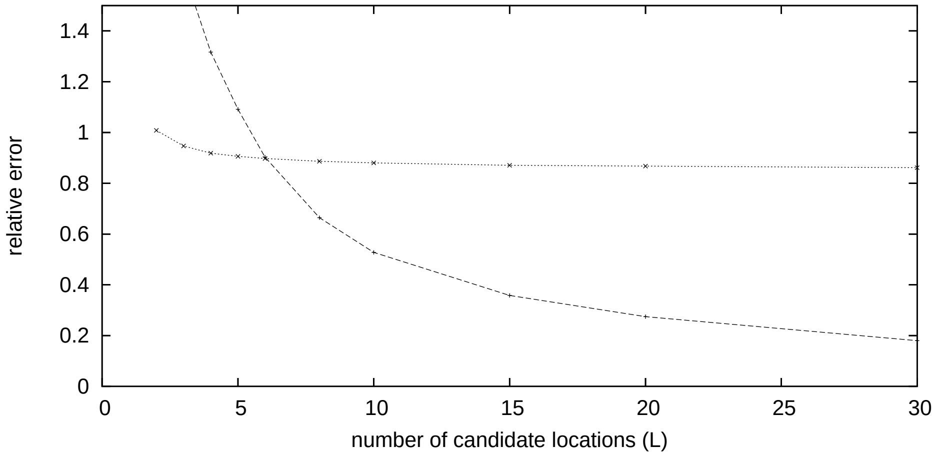

Fig. 4. Ratio of the average squared prediction error and average squared saccade length for the saccade predictor (dotted) and "ideal" predictor (dashed). The horizontal axis plots L, the number of salient locations per frame.

averaged over all video sequences and all test subjects.

For $L = 3$ and above, the saccade predictor achieves a relative error of less than 1, showing that, on average, the prediction moved in the right direction relative to the saccade starting point. The error decreases with increasing L, though beyond about $L = 10$ the improvement is small.

The relative errors achieved by the "ideal" predictor start out quite large for small L; initially, they are substantially larger than the error for the saccade predictor. This is explained by the fact that the saccade predictor makes its prediction relative to the starting point of the saccade; by limiting the size of this step, the saccade predictor can limit the size of the error relative to the total saccade length. The "ideal" predictor, on the other hand, only has the option of choosing one of the candidate locations; if all candidate locations are relatively far from the actual saccade target, the error made by the predictor will be relatively large.

For increasing L, the error for the "ideal" predictor decreases continuously, reaching about 0.2 for $L = 30$ candidate locations. This shows the potential for improvement that exists with the given saliency information.

The relative errors for the random and "maximum saliency" predictors are constant for all L: 4.51 for the random predictor and 3.49 for the "maximum saliency" predictor. (These values were not shown in the plot because they are rather large and would compress the scale of the vertical axis too much). This underlines the result from the previous section: A single salient candidate location does not capture the variance among individuals well and performs only slightly better than a location chosen at random.

4 Discussion and Outlook

The question of just how eye movements are triggered and controlled is far from being solved. Consequently, the problem of predicting eye movements remains a difficult one. Our results for the empirical saliency measure show that predicting the target of a saccade accurately from a saliency map alone is not possible. This is to be expected, given the influence of top-down attentional mechanisms – which are not based primarily on low-level image properties – on the scan-path. However, it seems feasible to select a small number of candidate locations, with a high probability that one of them will be chosen as the saccade target.

We have shown that saliency measures based on intrinsic dimensionality generate suitable candidate locations, though the results obtained using an "optimal" saliency measure (determined empirically from subjects' actual eye movements) demonstrate that there is still substantial room for improvement.

To date, we have not compared the performance of our saliency maps with other approaches proposed in the literature, though we plan to do so. Many of these existing approaches [5,6,11,12] are designed for static scenes; they could also be used for dynamic scenes on a frame-by-frame basis but would not be able to differentiate between static and moving objects. Since movement is one of the strongest bottom-up factors that influence eye movements, this is an obvious disadvantage. However, it appears that some of the existing saliency algorithms for static scenes (e.g. [5]) can relatively easily be generalized to dynamic scenes.

Some saliency algorithms for dynamic scenes exist in the literature [7,9]. How our algorithm fares compared to these other approaches remains to be seen. From a theoretical point of view, we find our algorithm appealing because of its conceptual simplicity. While other approaches employ features that are specifically dedicated to dynamic effects (motion, flicker etc.), our approach detects statically and dynamically salient regions with a single mechanism. However, we also point out that our algorithm has one major shortcoming in that it is based only on image intensity (luminance) and thus does not take colour differences into account, unlike other approaches. This is a limitation that we plan to address in the future.

For some applications, generating a small list of candidate saccade targets is sufficient. For instance, our work is motivated by applications involving the guidance of eye movements [10]. In this setting, it is sufficient to know a small number of locations that are candidates for the saccade target; the visual stimuli at one of these locations can then be enhanced while the others are suppressed so that the low-level image properties favour a saccade to the desired target.

We have also attempted to predict a single saccade target from the list of candidate locations, but our results have so far been quite modest. The average error is still about 0.86 of the average saccade length, even for large numbers of candidate locations.

This saccade predictor can certainly be improved upon. The current implementation 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. Also, our saccade predictor does not take into account the history of previously attended locations, which would be necessary to model higher-level phenomena, such as inhibition of return (a bias that tends to inhibit saccades to recently attended locations [24]).

Still, predicting a single saccade target well from only the visual input is a formidable task. It would involve modelling the complex top-down mechanisms that, as we speculate, select one saccade target from the list of candidate locations generated from low-level image properties. An important point that we make here, though, is that this list is rather short. Furthermore, one should not overestimate the unpredictability of eye movements: when predicting gaze while measuring it, the errors are rather small on average and limited to brief onsets of saccades [8,16]; in other words, if one displays a movie with true and predicted gaze overlayed, one can hardly see the difference (that becomes smaller, the higher the sampling rate is).

In summary, we have shown that saliency maps based on intrinsic dimensionality can be used to generate a short list of candidate locations that is likely to contain the actual saccade target. We plan to use this result in active displays to influence which of the candidates is actually chosen.

Acknowledgements

Information technology for active perception (Itap) [10] research is supported by the German Ministry of Education and Research (BMBF) as part of the interdisciplinary project *ModKog* (grant number 01IBC01B). We thank Karl Gegenfurtner and Jan Drewes of the Institute for General Psychology of the University of Gießen for acquiring the eye-tracking data first presented in [25]. We also thank the anonymous reviewers for their helpful comments.

References

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

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

[3] J. K. O' Regan, A. Noë, A sensorimotor account of vision and visual consciousness, Behavioral and Brain Sciences 24 (5) (2001) 939–1011.

[4] R. J. Leigh, D. S. Zee (Eds.), The Neurology of Eye Movements, Oxford University Press, 1999.

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

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

[7] G. Boccignone, A. Marcelli, G. Somma, Analysis of dynamic scenes based on visual attention, in: Proceedings of AIIA 2002, Siena, Italy, 2002.

[8] E. Barth, J. Drewes, T. Martinetz, Dynamic predictions of tracked gaze, in: Seventh International Symposium on Signal Processing and its Applications, Paris, 2003, special Session on Foveated Vision in Image and Video Processing.

[9] L. Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes, Visual Cognition (2005) (in press).

[10] Information technology for active perception. (2002).
URL http://www.inb.uni-luebeck.de/Itap/

[11] F. Stentiford, An estimator for visual attention through competitive novelty with application to image compression, in: Picture Coding Symposium, Seoul, Korea, 2001, pp. 101–104.

[12] T. Kadir, A. Zisserman, M. Brady, An affine invariant salient region detector, in: 8th European Conference on Computer Vision, Vol. 1, Springer, 2004, pp. 257–269.

[13] L. Itti, Models of bottom-up attention and saliency, in: L. Itti, G. Rees, J. K. Tsotsos (Eds.), Neurobiology of Attention, Elsevier, San Diego, CA, 2005, pp. 576–582.

[14] W. Becker, Saccades, in: R. H. S. Carpenter (Ed.), Vision & Visual Dysfunction Vol 8: Eye Movements, CRC Press, 1991, pp. 95–137.

[15] H. I. Christensen, J. Horstmann, T. Rasmussen, A control theoretical approach to active vision, in: Asian Conference on Computer Vision, 1995, pp. 201–210.

[16] E. Barth, J. Drewes, T. Martinetz, Individual predictions of eye-movements with dynamic scenes, in: B. Rogowitz, T. Pappas (Eds.), Electronic Imaging 2003, Vol. 5007, SPIE, 2003.

[17] M. Böhme, C. Krause, T. Martinetz, E. Barth, Saliency extraction for gaze-contingent displays, in: Proceedings of the 34th GI-Jahrestagung, Vol. 2, 2004, pp. 646–650, workshop on Organic Computing.

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

[19] C. Zetzsche, K. Schill, H. Deubel, G. Krieger, E. Umkehrer, S. Beinlich, Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach, in: P. R. et al. (Ed.), From Animals to Animats 5: Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior, Vol. 5, MIT Press, Cambridge, 1998, pp. 120–126.

[20] C. Mota, E. Barth, On the uniqueness of curvature features, in: G. Baratoff, H. Neumann (Eds.), Dynamische Perzeption, Vol. 9 of Proceedings in Artificial Intelligence, Infix Verlag, Köln, 2000, pp. 175–178.

[21] E. Barth, A. B. Watson, A geometric framework for nonlinear visual coding, Optics Express 7 (2000) 155–185.

[22] B. Jaehne, H. Haußecker, P. Geißler (Eds.), Handbook of Computer Vision and Applications, Academic Press, 1999.

[23] M. Böhme, C. Krause, E. Barth, T. Martinetz, Eye movement predictions enhanced by saccade detection, in: Brain Inspired Cognitive Systems, Stirling, United Kingdom, 2004.

[24] R. M. Klein, Inhibition of return, Trends in Cognitive Sciences 4 (2000) 138–147.

[25] M. Dorr, M. Böhme, J. Drewes, K. R. Gegenfurtner, E. Barth, Variability of eye movements on high-resolution natural videos, in: H. H. Bülthoff, H. A. Mallot, R. Ulrich, F. A. Wichmann (Eds.), Proceedings of the 8th Tübinger Perception Conference, 2005, p. 162.