DYNAMIC PREDICTIONS OF TRACKED GAZE

Erhardt Barth, Jan Drewes, and Thomas Martinetz

Institute for Neuro- and Bioinformatics, Ratzeburger Alle 160, D-23538 Luebeck, Germany

ABSTRACT

We present a model for predicting the eye-movements of observers who is viewing dynamic sequences of images. As an indicator for the degree of saliency we evaluate an invariant of the spatio-temporal structure tensor that indicates an intrinsic dimension of at least two. The saliency is used to derive a list of candidate locations. Out of this list, the currently attended location is selected according to a mapping found by supervised learning. The true locations used for learning are obtained with an eye-tracker. In addition to the saliency-based candidates, the selection algorithm uses a limited history of locations attended in the past. The mapping is linear and can thus be quickly adapted to the individual observer. The mapping is optimal in the sense that it is obtained by minimizing, by gradient descent, the overall quadratic difference between the predicted and the actually attended location.

1 INTRODUCTION

Vision is an active and highly selective process [1,2,3]. Therefore the message that is conveyed by an image depends very much on the scan-path, i.e., the sequence of eye movements that are used to look at an image. Visual communication systems, however, are based on only the classical image attributes, luminance and colour. In order to become part of visual communication systems, the scan-path should be sensed, processed, and "displayed" as suggested in [4,5]. For this purpose, our model shall help to better understand eye movements and make them more predictable. The foveated and active nature of vision has been used to derive a method of video compression [6] and thereby provided means of reducing the frustration one must feel when spending bits for transmitting visual information that nobody is looking at. Such methods involve the sensing of gaze direction but not the "display", i.e. they do not attempt to change the scan path.

The model presented here predicts gaze position at the current frame based on *previously attended locations* (known up to the previous frame) and *salient spatio-temporal features* (known at the current frame but derived from the current and previous frames). Models for eye-movements typically deal with static images and seem to converge on using a saliency map that models the bottom-up aspects of attention [7, 8]. Only few authors have considered dynamic scenes [9, 10, 11]. One major difficulty in modelling the top-down aspects is due to large inter-subject variances, i.e., observers use, to a certain extent, individual strategies for directing their gaze and attention. Therefore, in our approach we provide means of adapting the model to a particular observer. We also believe that top-down influences and also random components of the scan-path, are more significant when observers scan a static image for a longer period of time and less problematic with dynamic input.

2 THE MODEL

2.1 Temporal predictions

We know, by the use of an eye-tracker, a history of N locations attended in the past and want to predict the location that will be attended in the current frame at time t. The predicted location at time t is defined by:

$$X_t = X_{t-1} + A_{t-1} P_{t-1} \qquad (1)$$

where $X_t = (x_t, y_t)$ is the location predicted for the current frame and X_{t-1} is the previous location.

$P_{t-1} = (X_{t-2} - X_{t-1}, X_{t-3} - X_{t-1}, ..., X_{t-N} - X_{t-1})^T$ is an array of position vectors that holds the history of locations attended in the past. These locations are all expressed relative to the last currently known location X_{t-1}. The *Nx2* matrix P_{t-1} is mapped by the 1xN matrix A_{t-1} to a displacement vector that defines the shift of attention from the previous to the current frame. The matrix A_{t-1} is determined by supervised learning and is updated continuously.

2.2 The learning procedure

The learning rule is incremental and minimizes by gradient descent the mean prediction error, which is defined as the sum of quadratic differences between the predicted and the actually attended locations, i.e.

$$E = \frac{1}{t-1} \sum_{i=2}^{t} (X_{i-1} - X_{i-2} - A_{t-1} \cdot P_{i-2})^2 .$$

This error can be minimized by an iterative procedure, i.e. an incremental learning strategy, by using the following update rule [12]:

$$A_{t-1} = A_{t-2} + \varepsilon e P_{t-2}^T$$

where ε is the learning rate and $e = X_{t-1} - X_{t-2} - A_{t-2} \cdot P_{t-2}$ the current error that is used for incremental learning. The learning rate is the

distance by which the algorithm walks down the error function in the direction of the gradient $e \cdot P_{t-2}^T$. The learning rate we use is defined as

$$\varepsilon = \alpha \frac{eP^T Pe^T}{\left|P^T Pe^T\right|^2}.$$

The above expression is derived by first using a line-search method that minimizes the error on the current input and then scaling the result by α.

2.3 Predictions based on salient features

The above prediction based on previous locations cannot predict sudden shifts in attention that are due to, for example, the appearance of a novel object. We therefore extend the model to include predictions based on the actual input. In this extended model the predicted location is defined as:

$$X_t = X_{t-1} + A_{t-1} \cdot P_{t-1} + B_{t-1} \cdot S_t. \qquad (2)$$

The array $S_t = (..., X^C_{t,i} - X_{t-1,i}, ...)^T$ holds the salient candidate locations that are extracted from the current frame at time t. The index *i=1,...,L* denotes a number of up to L salient locations. The procedure for obtaining the salient locations based on a spatio-temporal saliency measure is described below. The *1xL* array B_{t-1} maps all the salient locations to a displacement vector that defines the saliency-based contribution to the shift of attention from the previous to the current frame. The actual shift is the sum of the saliency-based and the temporal contribution to the prediction. We obtain the matrix B_{t-1} by using the same learning procedure as for the matrix A_{t-1}. Note, however, that the matrices A_{t-1} and B_{t-1} are now learned simultaneously, i.e., the prediction error used to drive the learning procedure is obtained from predictions that involve both matrices. Of course, one could use only one matrix for both the previously attended and the saliency-based locations and only one matrix for the mapping. Nevertheless, the separation seems useful for conceptual reasons. Intuitively, the matrix A_{t-1} would learn to track and make short-time predictions, and the matrix B_{t-1} would rather learn a strategy for choosing a location from a list of candidate locations.

Attention is less likely to be directed towards a region of uniform intensity that does not change in time. In other words, the system is sensitive to changes. But what type of changes? In our view, the most basic categorization of changes is based on the concept of intrinsic dimensionality that has been introduced for images in [13] and shown to be useful for modelling attention with static images [14].

The intrinsic dimension of a 3-dimensional signal *f(x,y,t) is 0* if the signal is constant in all directions (*f(x,y,t)=c*), it is 1 if the signal is constant in 2 directions (*f(x,y,t)=g(u)*), it is 2 if the signal is constant in one direction (*f(x,y,t)=g(u,v)*), and it is 3 if there is no constant direction. Of particular relevance for the present context is the fact that i2D regions of images and image sequences, i.e. those image regions where the intrinsic dimension is at least 2, have been shown to be unique, i.e. they fully specify the image [15].

The evaluation of the intrinsic dimension of image sequences is possible within a geometric approach [16] and is here implemented by using the structure tensor **J**, which is well known in the computer-vision literature, see e.g. [17].

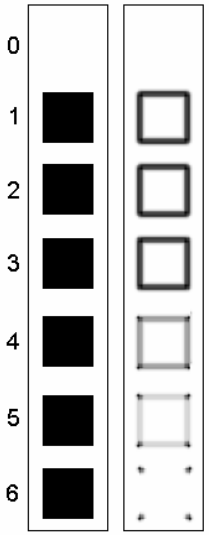
Based on the image-intensity function f(x,y,t), the structure tensor **J** is defined as:

$$J = w * \begin{pmatrix} f_x^2 & f_x f_y & f_x f_t \\ f_x f_y & f_y^2 & f_y f_t \\ f_x f_t & f_y f_t & f_t^2 \end{pmatrix}, \qquad (2)$$

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 zero if the eigenvalues of **J** are all zero, and in general it is *n* if *n* eigenvalues are different from zero. We use the invariant

$$S = M_{11} + M_{22} + M_{33}$$

of **J**, where M_{ij} are the minors of **J** obtained by eliminating the row 4-i and the column 4-j of **J**. The intrinsic dimension of f is at least 2 if S differs from zero. Therefore, the invariant S indicates non-redundant dynamic features. **Figure 1** on the left shows the response of S (right column) to an appearing square (left, time from top to bottom).



The $S(x, y, t)$ values are used to obtain a list of candidate locations as follows. Regions with $S(x, y, t) \le \theta$ are ignored. The threshold value θ remains a parameter of the model. Connected regions with S values above the threshold are reduced to only one location. This location is written to a ordered list $X^C_i = (x^C_i, y^C_i)$ of candidate locations with $i = 1, ..., L$. The list is ordered by the maximum and the mean values of S in the region. We have chosen this somewhat ad-hoc procedure to simplify the subsequent learning procedure.

2.4 The algorithm

To summarize, our algorithm involves the following steps:

1. Computation of the structure tensor **J** that is built from blurred products of first-order derivatives

(but can be built with more general linear filters as shown in [1818], i.e. with V1-like filters also). The derivatives have been estimated by discrete differences after low-pass filtering with a Gaussian kernel.

2. Computation of the invariant S(x,y,t) of **J**. S is estimated on multiple scales and the lower scales are sub-sampled to yield a pyramid.
3. Build of a list of L candidate locations based on S. The list is build by thresholding S, determining connected components, and choosing the location with maximum S as a candidate location.
4. The candidate list and N previously attended locations enter the incremental learning procedure that defines the matrices **A** and **B** of Equation (2).
5. The location predicted for the current frame is determined according to Equation (2).

3 PREDICTION RESULTS

For performance evaluation, we tested the model with our own recordings of eye-movements. The procedure involves four steps: (i) subjects watch a video sequence on a computer monitor and we record the direction of their gaze, (ii) we compute the saliency and the candidate locations based on the same video data, and (iii) we use the positions recorded for up to frame t-1 and the candidate locations based on saliency computed from up to frame t to make a prediction, and (iv) we use the difference between the prediction that we make and the actually attended location at frame t to train the matrices **A** and **B**. Hopefully, the errors will become smaller as we update the matrices.

We present results in terms of the prediction errors. The errors are compared among three different models. The first reference model (M1) is making predictions defined by $X_t = X_{t-1}$, i. e., we assume that gaze direction does not change between two frames.

The second model (M2) is making temporal predictions according to Equation (1) and the third model (M3) both temporal and saliency-based predictions according to Equation (2). The current parameters of the model are the following. The derivatives have been computed by finite differences after spatio-temporal Gaussian low-pass filtering with a kernel of size 5x5x5 and $\sigma_1 = 3$ for all variables *(x,y,t)*. The kernel w that convolves the product terms of the structure tensor **J** was the same as the one used for estimating the derivatives ($\sigma_2 = 3$). The influence of σ_1 and σ_2 on the prediction errors has not been analysed yet. The threshold θ was adaptive and set to 0.5 times the maximum of the current frame. For the results presented here we used a minimal configuration with *N=2* and *L=4*. The four saliency locations have been obtained by choosing only one location from each scale of the S pyramid. The learning rate had $\alpha = 0,001$ for the matrix **A** and $\alpha = 0,01$ for the matrix **B**.

3.1 Video sequences

Data were recorded for a synthetic and a natural video. The synthetic video, 204 frames long, was first showing a stationary square placed top right that disappeared at frame 45. At frame 60 a square appeared top left and moved to bottom right until it disappeared at frame 125. At frame 145 a square appeared gradually, stayed for 6 frames, and then disappeared gradually until frame 200. The real-life video was 735 frames long and showing a typical traffic scene. The size of the frames was 360 by 288 pixels, scaled to 800 by 600 pixels for full-screen playback. The sequences have been displayed on a 75 Hz computer monitor with a frame-rate of 25 frames per second, an image size of 40 times 30 cm at a viewing distance of 75 cm, thus spanning an horizontal field of view of about 30 deg. The movies will be made available on the institute's homepage.

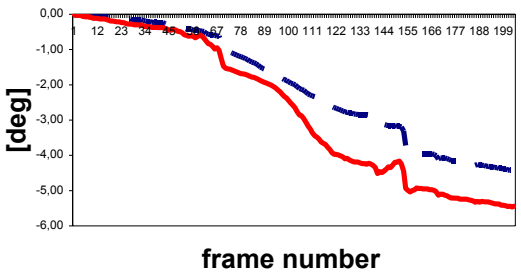


Figure 2. Prediction gain obtained for a synthetic scene. We show differences in cumulated absolute errors for model M2 versus M1 (dashed line) and model M3 versus M1 in deg of visual angle – see text for details.

3.2 Eye-movement recordings

Eye movements have been recorded by using the commercial system iViewX produced by the Sensomotoric Instruments GmbH. The eye tracker points a video camera to the observer's eye and uses two sources of infrared illumination to create two corneal reflexes that are tracked together with the pupil. The eye-tracker has been synchronized with the video sequences by our display program that was programmed to send a signal to the eye-tracker via the parallel port. The video display and the tracker were running on two different personal computers.

The prediction errors as a function of time are shown in Figure 2 for the synthetic sequence. We show the errors obtained with model M1 versus those obtained with model M2 (dashed line). Note that the predictions based on matrix **A** improve the prediction relative to the reference model M1. Note that an even greater improvement can be

obtained with model M3 that includes saliency-based prediction. Overall the improvements seem rather small but note that they have to be related to the sampling frequency of 25Hz and that we only show the improvements relative to model M1. For the traffic scene we have obtained a significant prediction gain for model M2 (cumulated gain of about 20 deg) but could not, at this size of the model, obtain any significant improvement for M3 relative to M2.

4 SUMMARY AND DISCUSSION

We have presented a model that can predict eye movements based on the history of previously attended locations and a saliency measure. Our approach differs from standard approaches in a number of ways. First, we deal with dynamic scenes and track the direction of gaze. Second, the model is partially derived from the current data by *machine-learning* techniques and can thus be *adapted to a particular observer* and used in real-time applications. Furthermore, the model can help to study eye-movements since it provides an *objective measure for* the *saliency* of a given image feature in terms of a prediction gain.
The performance of models of eye movements are hard to compare and are widely of qualitative nature in the literature. By relating the prediction gain due to a saliency measure to the prediction gain of optimal temporal predictions we have defined an objective measure for the prediction gain of a particular saliency measure. Nevertheless, comparison remains difficult because the prediction gain also depends on properties of the eye-tracker and the nature of the video data. However, for a given observer, given input, and eye-tracker the model will deliver a comparable prediction gain for both temporal and saliency-based predictions.

Our model is currently simple and small, as is the prediction gain that we obtain. However, various extensions are possible, e.g. a nonlinear coupling of the linear mappings **A** and **B** or even more complex, nonlinear mappings.

5 ACKNOWLEDGEMENTS

We thank Manuel Wille for implementing the program that displays the video and synchronizes the eye-tracker. *R*esearch is supported by the German Ministry of Education and Research as part of the interdisciplinary project *ModKog*. We thank the SensoMotoric Instruments GmbH for their eye-tracking support.

6 REFERENCES

1. D.M. McKay. Behind the eye. Basil Blackwell, 1991.
2. O Regan, A. Noe. A sensorimotor account of vision and visual consciousness, Behavioral and Brain Sciences, 24(5), 2001.
3. D. Noton, L. Stark. Eye movements and visual perception. Scientific American, Vol. 224, No. 6, pp. 34-43, 1971.
4. E. Barth. Information technology for active perception: Itap. First GRP-Symposium, Sehen und Aufmerksamkeit im Alter, Benediktbeuren, December, 2001.
5. E. Barth, T. Martinetz. Information technology for active perception. 8th Annual German-American Beckman Frontiers of Science Symposium, June, 2002.
6. Z. Wang, A. Bovik. Embedded Foveation Image Coding. IEEE TRANSACTIONS ON IMAGE PROCESSING, Vol. 10, No. 10, 2001.
7. I. A. Ryback, et.al. A model of attention-guided visual perception and recognition. Vision Research, 38, 2387-2400, 1998.
8. L. Itti, C. Koch, Computational Modelling of Visual Attention, Nature Reviews Neuroscience, Vol. 2, No. 3, pp. 194-203, Mar 2001.
9. C. M. Privitera, L. W. Stark. Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations, IEEE Trans. on PAMI, vol. 22, no. 9, pp. 970, 2000.
10. R. Milanese, S. Gil, T. Pun, "Attentive mechanisms for dynamic and static scene analysis", Optical Engineering, 34, No 8, pp. 2420-2434, 1995.
11. G. Boccignone, A. Marcelli, G. Somma. Analysis of dynamic scenes based on visual attention. Proceedings of AIIA 2002, Siena, Italy, September 2002.
12. S. Haykin. Neural Networks, 2d edition. Prentice-Hall, 1998.
13. C Zetzsche, 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, S. Beinlich. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. Proc. 5th Int. Conf. Simulation Adaptive Behav. 5, 120–126 (1998).
15. C. Mota, E. Barth. On the uniqueness of curvature features. Proceedings in Artificial Intelligence, Vol. 9: 175-178, 2000.
16. E. Barth, A. Watson. A geometric framework for nonlinear visual coding. Optics Express. Vol. 7. No. 4, 155-165, 2000. http://www.opticsexpress.org/oearchive/source/23045.htm.
17. B. Jaehne, H. Haußecker, P. Geißler, Eds., *Handbook of Computer Vision and Applications*, Academic Press, Boston, 1999.
18. C. Mota, I. Stuke, E. Barth. Analytic solutions for multiple motions. Proceedings of the 2001 International Conference on Image Processing, Thessaloniki, 2001.

.