Proceedings of the IS&T SPIE Annual Symposium on Electronic Imaging, Vol. 5007, Santa Clara, January 2003.

Individual predictions of eye-movements with dynamic scenes

Erhardt Barth[1], Jan Drewes, and Thomas Martinetz

Institute for Neuro- and Bioinformatics, University of Luebeck, Germany

ABSTRACT

We present a model that predicts saccadic eye-movements and can be tuned to a particular human observer who is viewing a dynamic sequence of images. 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 of adapting the model to a particular observer. As an indicator for the degree of saliency we evaluate the intrinsic dimension of the image sequence within a geometric approach implemented by using the structure tensor. Out of these candidate saliency-based locations, the currently attended location is selected according to a strategy found by supervised learning. The data are obtained with an eye-tracker and subjects who view video sequences. The selection algorithm receives candidate locations of current and past frames and a limited history of locations attended in the past. We use a linear mapping that is obtained by minimizing the quadratic difference between the predicted and the actually attended location by gradient descent. Being linear, the learned mapping can be quickly adapted to the individual observer.

Keywords: Eye-movements, saccades, saliency map, intrinsic dimension, machine learning, gaze-contingent display

INTRODUCTION

We deal with the problem of modelling eye-movements and attention. We approach this problem from a somewhat different perspective and therefore first motivate it by embedding it in the context of our research interests.

INFORMATION TECHNOLOGY FOR ACTIVE PERCEPTION

Our research is motivated by applications that involve gaze-contingent interactive displays and it is part of our research efforts that aim at *Information technology for active perception (Itap)*. As shown repeatedly, vision is a highly active process [1, 2, 3]. Therefore, the information conveyed by an image or a sequence of images is not only determined by light intensity and color, but to a large extent by the way in which a person is looking at an image. The sequence of eye-movements used to view a scene is called the scan-path [2]. The scan-path has been shown to be useful for object recognition [4] and video compression [5]. As a logically consistent extension of previous work the scan-path could be used to improve vision-based communication and behavior as proposed in [6, 7] . The idea is that the scan-path is an as important image attribute as are color and brightness and that this new image attribute needs to be *sensed* and *displayed*, and also processed, like the more traditional attributes are [6, 7].

For *sensing*, one can build on existing eye-tracking technology that needs to be improved, such that contact-free systems with sufficient spatial and temporal resolution can be built. This is still a challenging task but is becoming feasible due to new imaging techniques and increased computational power. The model we present in this paper shall contribute to this by enabling prediction-based eye-tracking. To "*display*" the new image attribute, one needs to guide the gaze of the viewing person according to a given scan-path. Techniques for such scan-path guidance will be obtained by closing and manipulating the loop between actively looking somewhere and reading out information there. The goal is to let people look at images and the environment with a given scan-path without them being aware of any manipulation. By enabling an analysis based on eye-tracking data with dynamic scenes, the model we present here can help to better understand how humans move their eyes and how eye movements could be guided. Regarding future perspectives, we believe that procedures for looking at images will become part of visual communication systems. Besides letting a person view images with the scan-path of another person, they may use a scan-path that could be either

[1] To whom correspondence should be sent. Email: barth@inb.uni-luebeck.de, Phone: +49 451 3909 583,
Address: Inst. f. Neuro- and Bioinformatics, Ratzeburger Alle 160, D-23538 Luebeck, Germany.

part of a knowledge base or computed in real time, such that, for example, a maximum gain of information can be achieved while looking at images and videos. An important area of applications is that of augmented vision with scan-path guidance as an interface. For example, in a car, the driver's attention can be directed to a pedestrian who has been detected by other sensors looking out of the car. We expect that *Itap* will have a large impact on visual communication. Current technical visual communication systems are based on the physical properties of images and they do not address the question of what message is conveyed by an image or a video. We believe that in the future communication systems, images and movies will be defined not only by brightness and color, but will be augmented with a recommendation of how to view the images, of what to see.

MODELS OF EYE MOVEMENTS AND ATTENTION

Models of visual attention typically deal with the scan-path that is used to view a static image – see [8] for a comprehensive review. The models seem to converge on using a saliency map that models the bottom-up aspects of attention. In most cases the saliency map is static. One major difficulty in modelling the top-down aspects is due to inter-subject variances, i.e. observers use, to a certain extent, individual strategies for directing their gaze and attention. For an in depth discussion see, for example, [9]. Because of the large inter-subject variances, 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. We avoid this problem by not letting it happen since the gaze-contingent display will be dynamic even with a static image. In most cases, however, we will deal with dynamic scenes. Although dynamic models of attention are of interest to robotics and active vision [10] only few authors have proposed models of attention with dynamic scenes [11, 12, 13] and we are not aware of a model that has been validated on a comprehensive set of dynamic and natural visual inputs. Partially this is because of a lack of data due to technological limitations.

PREDICTIONS BASED ON PREVIOUSLY ATTENDED LOCATIONS

As mentioned in the Introduction, our approach is motivated by gaze-contingent display. 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. The history of past locations can help to make this prediction and can be used for that scope in different ways. One way would be to mimic the computations in the vision system, for example by implementing inhibition of return. Another way would be to model the inertia of the moving eye, and the noise in the eye-tracker. We make no explicit attempt to differentiate these components but use supervised-learning techniques to find the best possible linear prediction one can make based on a history of size N. The predicted location is defined by:

$$X_t = X_{t-1} + A_{t-1} P_{t-1} \tag{1}$$

where $X_t = (x_t, y_t)$ is the vector-valued location predicted for the current frame and X_{t-1} is the previous location. $P_{t-1} = (X_{t-2} - X_{t-1}, X_{t-3} - X_{t-1}, ..., X_{n-N} - X_{n-1})^T$ 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.

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 [14]:

$$A_{t-1} = A_{t-2} + \varepsilon e P^T_{t-2}$$

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 $e \cdot P^T_{t-2}$. We have experimented with different constant learning rates and also with rates that have been decremented exponentially. Best results, however, have been obtained with a procedure that estimates the optimal learning rate at each iteration step and then weights this value with a constant value α. In this case the learning rate depends on the current error and is defined by

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

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

PREDICTIONS BASED ON SALIENT FEATURES

As shown in the Section Results, the above temporal predictions are good with respect to the mean prediction error. However, the prediction based on previous locations cannot predict sudden shifts in attention that are due to, for example, the sudden 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 $S_t = (X^C{}_{t,i} - X_{t-1}, X^C{}_{t-1,i} - X_{t-1}, ..., X^C{}_{t-M,i} - X_{t-1})^T$ holds a number of salient candidate locations that are extracted from the current frame at time t and also salient locations extracted from previous frames up to a history of *M* frames. The index *i=1,...,L* is used to denote a number of up to L salient locations estimated for every frame. Thus, S_t holds *M**L* locations in a *M**Lx2* matrix. The procedure for obtaining the salient locations is described below. The *1x M**L* matrix 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 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 and because we can update the matrices A_{t-1} and B_{t-1} at different rates. 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. It thus seems reasonable to use different learning rates.

INTRINSIC DIMENSION AND STRUCTURE TENSOR

The traditional view on attention modelling is that the system has a bottom-up and a top-down component. The bottom-up component drives the attention based on the image data and the top-down component is related to so called higher-level processes in the brain like, for example, long-term memory and motivation. The actually attended location would then be the result of the interplay between the two components. We also believe that eye-movements are, to a significant extent, driven by the visual input – see Discussion. 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? What we would like to have is a generic alphabet of change types with the different types having different amounts of saliency. Such an alphabet should classify a constant and static region with lowest saliency, stationary edges and uniform regions that change in time with intermediate, and popping regions that have spatial structure with high saliency. Although this issue has been analysed from different perspectives, we believe that, from a theoretical point of view, there exists a basic alphabet of change types that should be considered first and has not yet received adequate attention in the literature. Accordingly, our approach to saliency is based on the concept of intrinsic dimensionality that has been introduced for images in [15] and shown to be useful for modelling attention with static images [16]. Here, we use the intrinsic dimension of the image sequence as an indicator for the degree of saliency.

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 [17].

The evaluation of the intrinsic dimension is possible within a geometric approach [18] and is here implemented by using the structure tensor **J**, which is well known in the computer-vision literature (see e.g. [19]).

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

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

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. However, we do not perform the eigenvalue analysis of **J** since it is possible to derive the intrinsic dimension from the invariants of **J**, which are:

$$H = \frac{1}{3}\text{trace}(J) = \lambda_1 + \lambda_2 + \lambda_3$$

$$S = M_{11} + M_{22} + M_{33} = \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3$$

$$K = \det(J) = \lambda_1\lambda_2\lambda_3$$

where M_{ij} are the minors of **J** obtained by eliminating the row 4-i and the column 4-j of **J**. The λ_i are the eigenvalues of **J**. Note, however, that we do not need to estimate these eigenvalues. Since J is a positive definite matrix, the intrinsic dimension is at least 1 if H differs from zero, at least 2 if S differs from zero, and 3 if K differs from zero.

THE CANDIDATE LIST

Our current implementation is limited to using only the invariant S for saliency. This seems the simplest choice because a value of $S \neq 0$ indicates an intrinsic dimension of at least 2 and, therefore, suppresses regions of dimension less than 2, which are redundant. The $S(x, y, t)$ values are then used to obtain a list of candidate locations as follows. Regions with $S(x, y, t) \leq \theta$ 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.

THE ALGORITHM

To summarize, our algorithm involves the following steps:

1. Computation of the structure tensor J(x,y,t) that is built from blurred products of first-order derivatives (but can be built with more general linear filters as shown in [20], 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 built by thresholding S, determining connected components, and choosing the location with maximum S as a candidate location.
4. The candidate list with a history of M frames 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).

RESULTS

For performance evaluation, we trained and tested the model with our own recordings of eye-movements. We present results in terms of the prediction errors. The errors are compared for three different models. The first model (M1) is making predictions defined by

$$X_t = X_{t-1}.$$

The second model (M2) is making temporal predictions according to Equation (1) and the third model (M3) temporal and saliency-based predictions according to Equation (2).

The videos that show input sequences, saliency measures, predicted locations and actually attended locations will be presented at the conference and can be viewed on the web at http://www.inb.uni-luebeck.de/Itap/ . The 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*, *M=1* 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 was scaled by $\alpha = 0{,}001$.

VIDEO SEQUENCES

The results presented here have been obtained with two video sequences. The first one was synthetic, 750 frames long, and showing a square that moves from the top left to the bottom right. In addition, other two squares pop in and out at different moments. The second sequence was a real-life video, 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.

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.

PREDICTION ERRORS

The prediction errors as a function of time are shown in Figure 1 for the synthetic sequence. Figure 1a shows the shifts of gaze (squared vector lenght), Figure 1b the errors obtained with model M1 versus those obtained with model M2. Note that the temporal prediction improves the prediction mainly at gaze shifts. Figure 1c shows that an even greater improvement can be obtained with model M3 that included saliency-based prediction. Note the different scaling of the y-axis. Figure 2 shows the differences in cumulated errors, in (a) for model M2 versus M1, in (b) for model M3 versus M1 and in (c) for model M3 versus M2. Figure 3 shows the same results as Figure 2 but for the traffic scene. The parameters were the same for both movies, except for the number of frames and the fact that we started with different initialization values for the matrix **A**. Overall the improvements are rather small but encouraging, given the size of the model, and the small number of parameters that still need to be optimized.

DISCUSSION

Our primary current goal is to better understand human eye-movements of observers that view natural dynamic scenes. In a later stage the model will be used to optimise the design of gaze-contingent interactive displays. Different from many other models, our goal is to perform online predictions by knowing the locations attended in the past and computing salient features from the image stream. However, the model has a potential for analysing strategies of the eye-movement system and the saliency of different features. Results of such an analysis have not been reported here but will be important for determining features that can attract attention such as to allow for guidance of the eye movements.

We started with the simplest assumptions about saliency and the decision process. From the perspective of multidimensional signal processing, looking at intrinsic dimensions higher than one seems the natural extension of the one-dimensional concept of a temporal transient. Learning the optimal linear mapping from sample data also seems the most straightforward approach to the problem of determining the attended location given the visual input. Of course, vision is more complex, but more complex models must be related to such simple ones. However, our model can be extended in a number of ways. In our current implementation, we predict the displacement vector by searching for optimal weights that are then used to scale the length of previous displacement vectors. We could generalize the model such as to also rotate these vectors. Significantly more candidate locations could be used by including additional saliency measures. Furthermore, the linear mappings could be replaced by nonlinear mappings, e.g. neural networks. Nonlinearities will be needed to decouple the temporal tracking and the saliency-based shifts.

Many efforts have been made to understand the scan path of observers that view static images. Maybe the visual system has not been optimised for this situation and, therefore, random and top-down influences become more important. We believe that with a dynamic visual input the eye movements are more natural and easier to predict. Still, different observers may have very different scan paths and, therefore, see different things with the same visual input. For this reason we have designed a model that can be tuned to a particular observer. For the same reason we decided to not only observe the scan path but to change it such as to improve visual communication and vision-based interaction. Our model will help to do that.

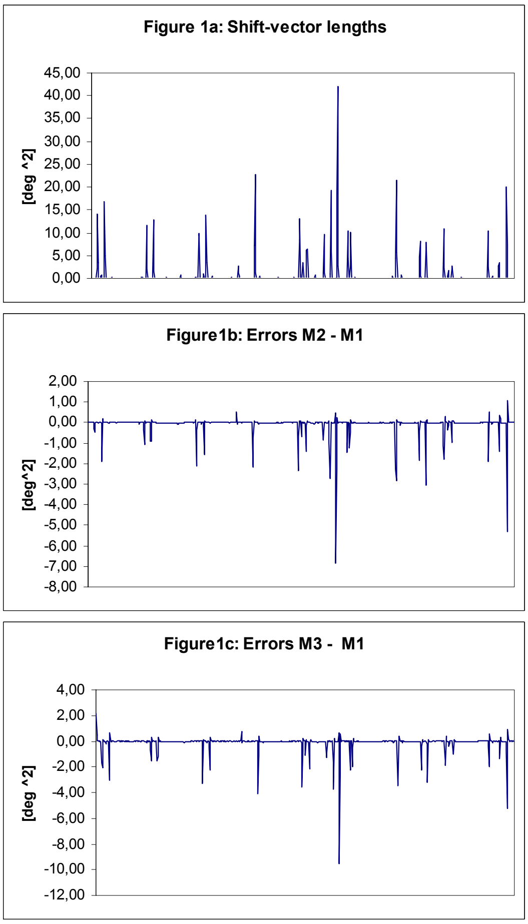


Figure 1: Prediction errors for synthetic scene.

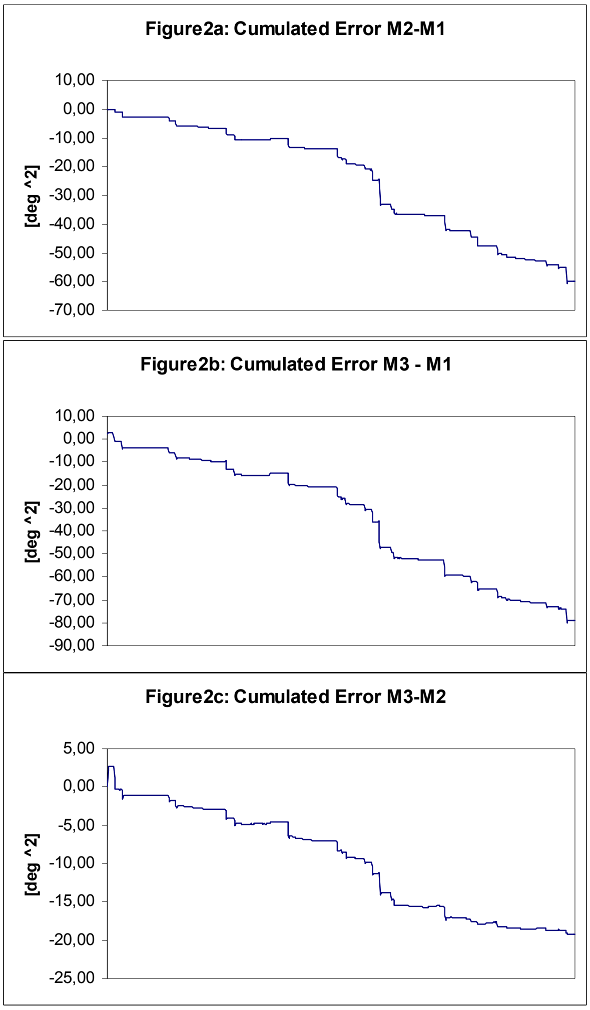
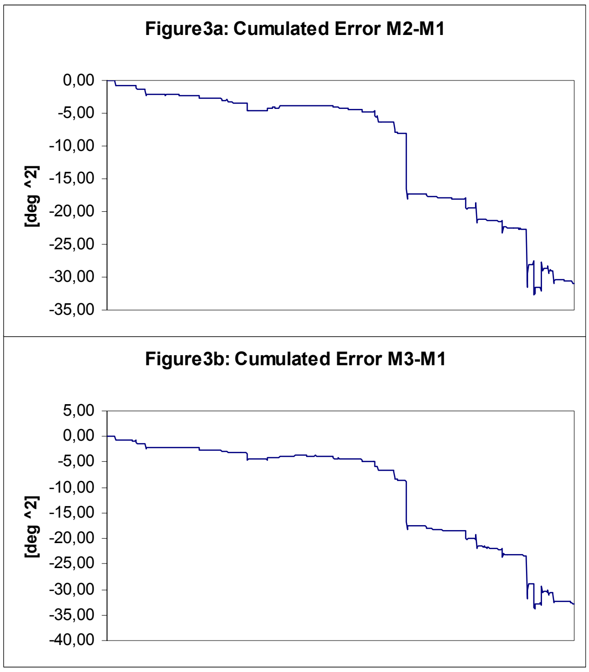


Figure 2.
Cumulative prediction errors for synthetic scene.



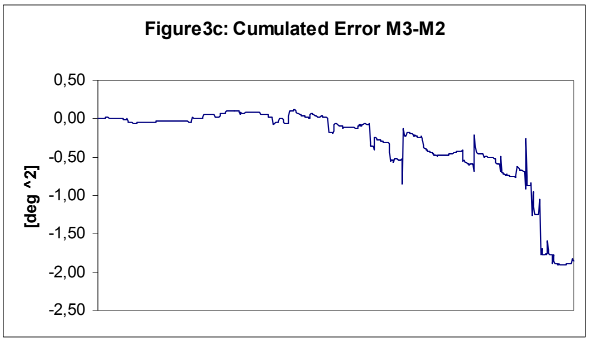


Figure 3:
Cumulative predictions for traffic scene.

ACKNOWLEDGEMENTS

We thank Manuel Wille for implementing the program that displays the video and synchronizes the eye-tracker. *Itap* research is supported by the German Ministry of Education and Research as part of the interdisciplinary project *ModKog*. We thank C. Mota and G. Boccignone for valuable suggestions and comments on the manuscript. We also thank the SensoMotoric Instruments GmbH for their eye-tracking support; data have been obtained with their iViewX system.

REFERENCES

1. D.M. McKay. Behind the eye. Basil Blackwell, 1991.
2. D. Noton, L. Stark. Eye movements and visual perception. Scientific American, Vol. 224, No. 6, pp. 34-43, 1971.
3. O Regan, A. Noe. A sensorimotor account of vision and visual consciousness, Behavioral and Brain Sciences, 24(5), 2001.
4. I. A. Ryback, et.al. A model of attention-guided visual perception and recognition. Vision Research, 38, 2387-2400, 1998.
5. Z. Wang, A. Bovik. Embedded Foveation Image Coding. IEEE TRANSACTIONS ON IMAGE PROCESSING, Vol. 10, No. 10, 2001.
6. E. Barth. Information technology for active perception: Itap. First GRP-Symposium, Sehen und Aufmerksamkeit im Alter, Benediktbeuren, December, 2001.
7. E. Barth, T. Martinetz. Information technology for active perception. 8th Annual German-American Beckman Frontiers of Science Symposium, June, 2002.
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. B. Adams, C. Breazeal, R. Brooks, B. Scassellati. Humanoid Robots: A New Kind of Tool, IEEE Intelligent Systems, Vol. 15, No. 4, July/August, 2000, pp. 25—31.
11. G. Giefing, H. Janßen, H. Mallot: Saccadic Object Recognition with an Active Vision System. Proceedings of the European Conference on Artificial Intelligence, 803-805, 1992.
12. R. Milanese, S. Gil, T. Pun, "Attentive mechanisms for dynamic and static scene analysis", Optical Engineering, 34, No 8, pp. 2420-2434, 1995.
13. G. Boccignone, A. Marcelli, G. Somma. Analysis of dynamic scenes based on visual attention. Proceedings of AIIA 2002, Siena, Italy, September 2002.
14. S. Haykin. Neural Networks, 2d edition. Prentice-Hall, 1998.
15. C Zetzsche, E Barth. Fundamental limits of linear filters in the visual processing of two-dimensional signals. Vision Research, 30:1111-1117,1990.
16. 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).
17. C. Mota, E. Barth. On the uniqueness of curvature features. Proceedings in Artificial Intelligence, Vol. 9: 175-178, 2000.
18. 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.
19. B. Jaehne, H. Haußecker, P. Geißler, Eds., *Handbook of Computer Vision and Applications*, Academic Press, Boston, 1999.
20. C. Mota, I. Stuke, E. Barth. Analytic solutions for multiple motions. Proceedings of the 2001 International Conference on Image Processing, Thessaloniki, 2001.

Note: The movies used to train the model, prediction results, and some of the cited references can be downloaded from our homepage http://www.inb.uni-luebeck.de/ .