Guiding the mind's eye: improving communication and vision by external control of the scanpath

Invited contribution for a special session on Eye Movements, Visual Search, and Attention: a Tribute to Larry Stark

Erhardt Barth[a], Michael Dorr[a], Martin Böhme[a], Karl Gegenfurtner[b], and Thomas Martinetz[a]

[a]Institute for Neuro- and Bioinformatics, University of Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany

[b] Allgemeine Psychologie, Justus-Liebig-University, Otto-Behaghel-Str. 10, D-35394 Gießen, Germany

ABSTRACT

Larry Stark has emphasised that what we visually perceive is very much determined by the scanpath, i.e. the pattern of eye movements.[1] Inspired by his view, we have studied the implications of the scanpath for visual communication and came up with the idea to not only sense and analyse eye movements, but also guide them by using a special kind of gaze-contingent information display. Our goal is to integrate gaze into visual communication systems by measuring and guiding eye movements. For guidance, we first predict a set of about 10 salient locations. We then change the probability for one of these candidates to be attended: for one candidate the probability is increased, for the others it is decreased. To increase saliency, for example, we add red dots that are displayed very briefly such that they are hardly perceived consciously. To decrease the probability, for example, we locally reduce the temporal frequency content. Again, if performed in a gaze-contingent fashion with low latencies, these manipulations remain unnoticed. Overall, the goal is to find the real-time video transformation minimising the difference between the actual and the desired scanpath without being obtrusive. Applications are in the area of vision-based communication (better control of what information is conveyed) and augmented vision and learning (guide a person's gaze by the gaze of an expert or a computer-vision system). We believe that our research is very much in the spirit of Larry Stark's views on visual perception and the close link between vision research and engineering.

Keywords: eye movements, scanpath, eye tracking, gaze-contingent display, gaze guidance, saliency, foveation, augmented vision

1. INTRODUCTION

One of the most important characteristics of human vision is that we must constantly shift our gaze between objects of interest because only a small part of the retina (the fovea) provides high visual acuity. Additionally, we can only attend to a very limited number of features and events in the visual environment. These facts have severe consequences for visual communication, because what is communicated depends to a large degree on those mechanisms in the brain that deploy our attentional resources and determine our eye movements.

The information that is conveyed by an image is thus determined not only by the image itself, but by the image in conjunction with the observer's gaze pattern, which may vary considerably from person to person and with context. Therefore gaze is as important an attribute as brightness or colour for defining the message that reaches the observer, but existing technology does not take this into account.

We propose that future information and communication systems should be designed to optimise gaze patterns and the use of the user's limited attentional resources. We believe that in future communication systems images and movies will be defined not only by brightness and colour, but will be augmented with a recommendation of where to look, of how to view the images.

Gaze guidance can also be used to create new kinds of vision aids that fuse the strengths of human and computer vision to improve the visual capabilities of the user. Such augmented-vision systems are of particular interest for automotive

Further author information: (Send correspondence to EB)
EB, MD, MB, TM: E-mail: barth, dorr, boehme, martinetz@inb.uni-luebeck.de
KG: Karl.R.Gegenfurtner@psychol.uni-giessen.de

Figure 1. Hardware currently used to incorporate eye tracking technology into visual communication systems. Left: Monocular head-mounted display worn together with a head-mounted EyeLink eye tracker. Middle: Remote eye tracker attached to a standard TFT screen. Right: Custom-built head-mounted display with integrated binocular eye tracking and two outward-facing scene cameras.

applications. For example, the driver's attention can be directed towards a pedestrian, who has been detected by sensors looking out of the car, in cases when the driver would otherwise fail to see the pedestrian.

Gaze guidance can be used for a further kind of application in which novices can be taught to view images with the eyes of experts. It is known that experts, for example experienced pilots, scan their environment in a way that substantially differs from how inexperienced viewers would. We believe that by applying the gaze pattern of experts to novices, we can evoke a sub-conscious learning effect.

To reach such goals, however, a considerable amount of basic research and technological development is still required. In Fig. 1 we show the eye-tracking and display technology that we currently use and further develop. In the remainder of this paper we report on a few results that address selected problems of analysing, predicting, and guiding gaze.

2. ANALYSIS OF GAZE

We have investigated the variability of eye movements with dynamic natural scenes. To this end, we collected a large data set of gaze samples from 54 subjects watching a variety of short high-resolution video clips (20 s duration each). For each movie frame, clusters of gaze samples were extracted by an unsupervised machine-learning algorithm. First, a fixation map was created by a superposition of Gaussians centred at each gaze sample. From the resulting map, up to $n = 20$ maxima were extracted by iteratively applying a lateral inhibition scheme. Then, clusters were formed using a simple distance threshold. Results show that there exist hot spots that contain a high number of fixation locations. On average, 5-15 clusters (2-5 % of the viewing area) account for 60 % of all fixations. An example of fixation patterns is shown for a single frame of one of our videos in Fig. 2. The data on eye-movement variability have been briefly presented at two conferences[2,3]; a comprehensive analysis will be published in a forthcoming paper.

3. PREDICTION OF GAZE

Our approach to gaze prediction divides the problem into two parts: First, predicting the eye movements made between saccades (intersaccadic prediction); and second, predicting the targets of saccades (saccade prediction). 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.[4,5]

Saccade prediction is certainly the harder of the two subproblems. Like other authors, e.g. Ref. 6, 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 (see Fig. 3 for an example). 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. 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

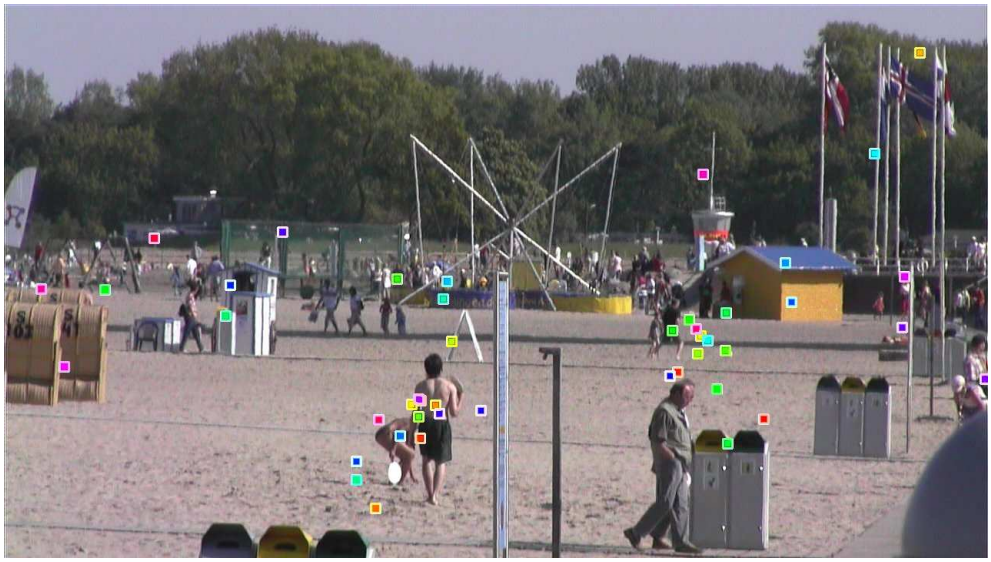

Figure 2. Still shot from a video of a natural scene. Each little square indicates the gaze position of one observer. These gaze positions are not uniformly distributed across the image, but tend to cluster in "interesting" regions.

Figure 3. Left: Still shot from a video. Right: Corresponding K saliency map. Ten candidate saccade locations were extracted from the map (little squares).

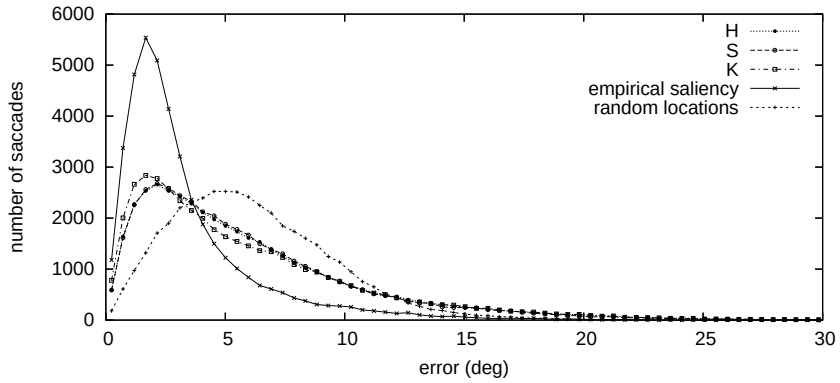

Figure 4. Saliency prediction results. Histogram of error (distance of saccade target to closest salient candidate location) for ten candidate locations. The horizontal axis plots the error magnitude in degrees, the vertical axis plots the number of saccades per histogram bin. Plots are shown for the H, S, and K saliency measures, the empirical saliency measure, and locations chosen at random. When using only one instead of ten candidates, errors based on all 4 saliency measures are still better than chance but unacceptably high.

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 ten, that will with high probability include the actual saccade target, and our above mentioned results on gaze analysis show that a small number of target locations usually covers most of the variations in the eye movements made by different observers.

As described previously,[4,5] our approach to saliency is based on the concept of intrinsic dimension.[7–9] 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.[10,11] The evaluation of the intrinsic dimension is possible within a geometric approach that is plausible for biological vision[9] and is implemented here by using the structure tensor $\mathbf{J}$, which is well known in the computer-vision literature.[12,13] Our saliency measures are the invariants of $\mathbf{J}$, namely the trace H, the sum of minors S, and the determinant K. 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.

As a baseline for assessing the saliency maps computed analytically, based on the H, S, and K measures 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 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 (the empirical saliency is actually doing even better than the best possible saliency because it has been derived from the data that are then predicted). 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 16 pixels at each of these positions. The superposition of these Gaussians then yields the empirical saliency map. For a detailed description, see Ref. 5.

4. GAZE GUIDANCE

The final goal of our gaze guidance system is to direct the user's attention to a specific part of a scene, ideally without the user noticing this guidance. Our strategy is to (i) predict a few candidate locations, (ii) increase the probability of being attended for one candidate, and (iii) decrease the probability for the other candidates. We have not yet implemented a system that integrates all components of this strategy, but will show some results related to (ii) and (iii) above in the remainder of this section.

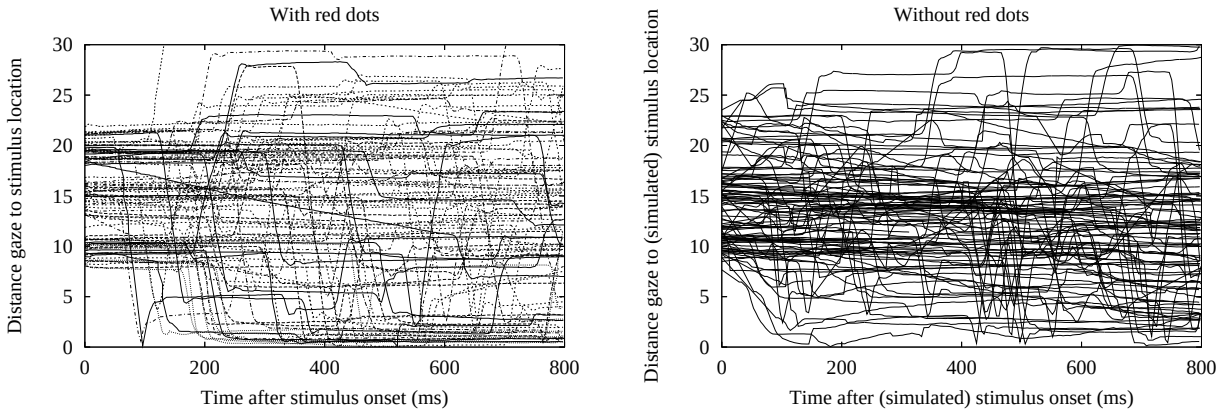

Figure 5. Raw data plots for stimulation with red dots. Horizontal axis plots time after stimulation onset, vertical axis denotes the distance from the gaze position to the locus of the stimulation. Left: Results for one subject. Right: Baseline trials where red dots were not shown (since red dots had been shown at randomly chosen *salient* locations this control seems necessary).

4.1. Effect of gaze-contingent red dots

Apart from our work on modelling which image features attract gaze, we have therefore also conducted experiments with several different spatio-temporal transformations designed to alter eye-movement characteristics. These transformations were based on observations made with synthetic stimuli, which are commonly used in experiments that investigate attentional effects. The first set of transformations was motivated by the well-known fact that sudden object onsets in the visual periphery can attract attention. We chose to briefly superimpose small bright red dots on the movie. Depending on the eccentricity of the dots at the time of their flashing, in up to about 40 % of trials, saccades were initiated towards the location of the flashed red dot when subjects where asked to just watch the movie. The results of these experiments are shown in Fig. 5.

Because the typical saccadic latency of about 200 ms exceeds the presentation time of the dot, which was set to 120 ms, the red dot was already switched off by the time the saccade was finished, so that in about 50 % of cases this stimulation remained invisible in another set of experiments where (other) subjects watched the same movies, but were now asked to detect the red dots, and they pressed a button when they detected the dots.

Similar effects were obtained in an experiment where the red dot was replaced by a looming stimulus, the looming stimulus being harder to detect than the red dot. Nevertheless, the exact parameters for an optimal guidance effect, such as size, contrast, duration, or the timing with regard to previous saccades, still need to be determined.

4.2. Effect of gaze-contingent spatio-temporal filtering

For a second, more complex set of transformations, we have developed a gaze-contingent display that can in real time change the spatio-temporal content of an image sequence as a function of where the observer is looking.[14]

Gaze-contingent displays manipulate some property of the (static or moving) image as a function of gaze direction (see Ref. 15 for a review). This type of display was first used in reading research[16] and has since been used in many psychophysical and perceptual studies (e.g. Ref. 17). The image property that is most commonly manipulated in a gaze-contingent display is spatial resolution. A popular type of manipulation is to foveate an image or video, i.e. to simulate the effect of the variable resolution of the human retina, which is highest at the fovea and falls off towards the periphery. If the foveation is adjusted to match the resolution distribution of the retina, the effect is not noticeable for the observer, but the resulting images can be compressed more efficiently because they contain less high-frequency content.[18,19] Another application is to visualise the effect of diseases of the eye, e.g. glaucoma[20]; these visualisations can be used to educate students or family members of patients about the effects of such diseases. The current state-of-the-art algorithm for gaze-contingent spatial filtering of video is due to Perry and Geisler.[20] Unlike previous algorithms, which introduced artifacts in the filtered images, their algorithm produces smooth, artifact-free results.

Figure 6. Left: Image from one of the video sequences. Right: Same image with gaze-contingent temporal filtering applied; the white square at the centre left (below the white sail) indicates the point of regard.

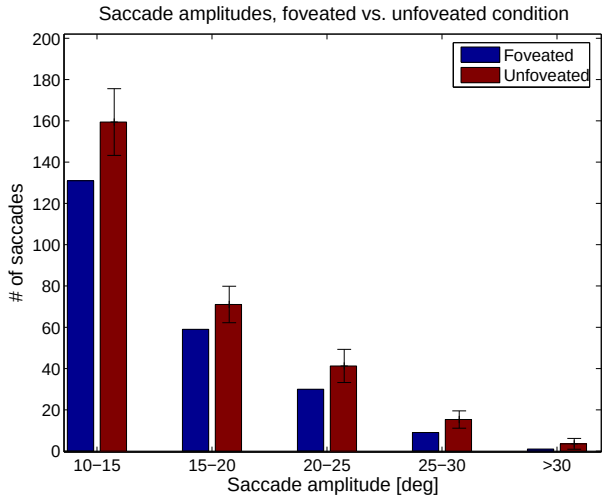

Figure 7. Histogram of saccade amplitudes with and without gaze-contingent temporal filtering. Only the histogram bars for saccades of 10 degrees and greater are shown since no effect could be expected in the central field of view due to the shape of the foveation function.

Based on this work, we have developed a gaze-contingent display that manipulates not the spatial, but the temporal resolution of a video. The basic effect of temporal filtering is to blur the moving parts of an image while leaving the static parts unchanged (see Fig. 6 for an example). Our motivation for performing this type of manipulation is that we want to examine the effect that it has on eye movements; movement or change in the periphery of the visual field is a strong cue for eye movements. The results presented in Fig. 7 show that gaze-contingent temporal filtering reduces the number of saccades with large amplitude.

To further improve the effect of the gaze-contingent display, we plan to specifically change the spatio-temporal content only at certain locations in an image.

4.3. Visibility of temporal blur on a gaze-contingent display

In the following, we will describe an experiment where we locally and selectively suppress temporal frequencies. The locus of the suppression is then varied so as to investigate the visibility of such changes as a function of eccentricity. As opposed to the temporal foveation described above, the video is now filtered with a spatial annulus with variable width and eccentricity. Note that we do not intend to measure the threshold for the maximum temporal frequency that can be detected at a given eccentricity; it is the absence of higher temporal frequencies that subjects had to detect.

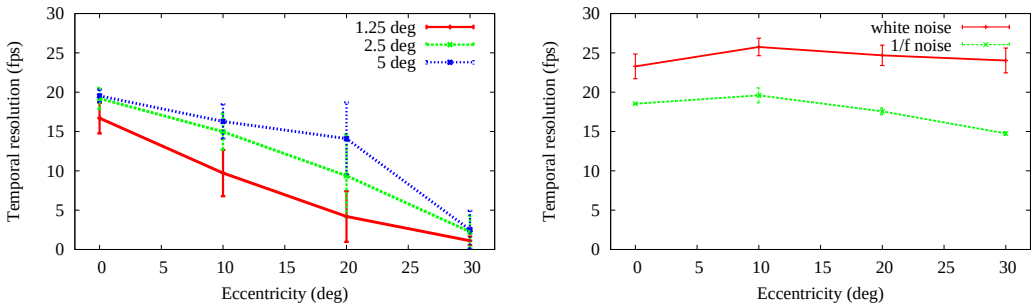

Figure 8. Detection thresholds for temporal filtering as a function of eccentricity. Left: Filtering applied to natural scenes. Each line shows results for one width of the filtered annulus-shaped region (as indicated in the legend). Right: Data for synthetic noise sequences with differing spectral characteristics; a single ring width of 2.5 degree has been used in these latter experiments.

We had shown previously that in natural scenes, a selective filtering of higher temporal frequencies in an annulus-shaped region around the centre of gaze remains unnoticed over a wide range of frequencies.[21] As shown in the left part of Fig. 8, the level to which temporal blur can be introduced increases dramatically with eccentricity.

We now compare these findings with the visibility of temporal blur in dynamic noise with varying spectral characteristics. We tested white noise (uniform spectral content) as well as pink noise (1/f spectral falloff) that was temporally lowpass-filtered to model the temporal correlation across consecutive frames that is characteristic for natural scenes. As in the case of the experiments conducted with natural images, three subjects watched 20 s long video clips (1024x576 pixels spatial, 30 frames per second temporal resolution) where higher temporal frequencies were filtered only in an annulus-shaped region of width 2.5 degrees at an eccentricity of 0, 10, 20, or 30 degrees. After stimulus presentation, subjects had to indicate whether they had perceived any temporal blur. Threshold frequencies were then adjusted in an interleaved staircase procedure. Subjects were seated 55 cm from a 22" CRT computer screen, running at a spatial resolution of 1280x960 pixels and a refresh rate of 90 fps. Eye movements were measured at 240 Hz and the latency of the system, defined as the time from a change in gaze position to the update of the display, was 35-75 ms (60 ms on average). This latency may seem quite high compared to the latencies of gaze-contingent displays that alter single images only, but note that building a temporal multi-resolution pyramid is by far costlier. Ultimately, every video frame that is displayed is a weighted sum of more than 250 high-resolution frames surrounding it.

The results in the right part of Fig. 8 clearly show that there is a qualitative difference between natural and synthetic sequences. Contrary to the effect found in natural scenes, detection thresholds for temporal blur do not vary significantly with eccentricity both in white and pink noise sequences. However, in white noise, temporal blur can be detected more easily (detection threshold at about 25Hz) than in pink noise (18Hz). We do not have an explanation yet for the qualitative differences described above.

5. CONCLUSIONS

We have shown that a rather small set of locations (10-20) where people may look while watching a natural video can be predicted with acceptable errors based on simple low-level dynamic saliency measures. This fits well with our analysis of eye movements on high resolution natural videos, which shows that eye movements tend to cluster in rather few locations (10-20). We have then reported on simple experiments that were meant to increase the saliency by the gaze-contingent and brief presentation of red dots in the periphery. Related to methods that would decrease saliency, we have shown that spatio-temporal blur changes the gaze pattern by inhibiting saccades. Since we are looking for unobtrusive ways of guiding gaze, we also analysed the visibility of temporal blur and found that peripheral temporal blur remains unnoticed on our gaze-contingent display, an effect that cannot be reproduced on the synthetic movies that we used.

We would therefore like to conclude that gaze guidance seems possible in principle, but more work is required to better understand the most efficient ways of performing it. Eventually we will have to show that gaze guidance can improve human vision capabilities in behavioural tasks and by that justify the applications that we have in mind.

ACKNOWLEDGMENTS

We dedicate this paper to Larry Stark who has influenced our way of thinking about visual perception. Research has been supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01B with acronym *ModKog*.

REFERENCES

1. *D. Noton and L. Stark, "Eye movements and visual perception,"* Scientific American **224***(6), pp. 34–43, 1971.*
2. *M. Dorr, M. Böhme, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Analysing and reducing the variability of gaze patterns on natural videos," in* Proceedings of 13th European Conference on Eye Movements, *M. Groner, R. Groner, R. Müri, K. Koga, S. Raess, and P. Sury, eds., p. 35, 2005.*
3. *M. Dorr, M. Böhme, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements on natural videos," in* Proceedings of the BIP Workshop on Bioinspired Information Processing, Lübeck, Germany, *2005.*
4. *E. Barth, J. Drewes, and T. Martinetz, "Individual predictions of eye-movements with dynamic scenes," in* Electronic Imaging 2003, *B. Rogowitz and T. Pappas, eds.,* **5007**, *pp. 252–259, SPIE, 2003.*
5. *M. Böhme, M. Dorr, C. Krause, T. Martinetz, and E. Barth, "Eye movement predictions on natural videos,"* Neurocomputing *, 2006. (in press).*
6. *L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis,"* IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, *pp. 1254–1259, Nov 1998.*
7. *C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two-dimensional signals,"* Vision Research **30**, *pp. 1111–1117, 1990.*
8. *C. Zetzsche and E. Barth, "Image surface predicates and the neural encoding of two-dimensional signal variation," in* Human Vision and Electronic Imaging: Models, Methods, and Applications, *B. Rogowitz, ed.,* **SPIE 1249**, *pp. 160–177, 1990.*
9. *E. Barth and A. B. Watson, "A geometric framework for nonlinear visual coding,"* Optics Express **7***(4), pp. 155–165, 2000.*
10. *C. Mota and E. Barth, "On the uniqueness of curvature features," in* Dynamische Perzeption, *G. Baratoff and H. Neumann, eds.,* Proceedings in Artificial Intelligence **9**, *pp. 175–178, Infix Verlag, (Köln), 2000.*
11. *E. Barth, T. Caelli, and C. Zetzsche, "Image encoding, labeling, and reconstruction from differential geometry,"* CVGIP: Graphical Model and Image Processing **55**, *pp. 428–46, November 1993.*
12. *B. Jaehne, H. Haußecker, and P. Geißler, eds.,* Handbook of Computer Vision and Applications, *Academic Press, 1999.*
13. *E. Barth, "The minors of the structure tensor," in* Mustererkennung 2000, *G. Sommer, ed., pp. 221–228, Springer, (Berlin), 2000.*
14. *M. Böhme, M. Dorr, T. Martinetz, and E. Barth, "Gaze-contingent temporal filtering of video," in* Eye Tracking Research and Applications (ETRA), *2006. (in press).*
15. *A. T. Duchowski, N. Cournia, and H. Murphy, "Gaze-contingent displays: A review,"* CyberPsychology & Behavior **7***(6), pp. 621–634, 2004.*
16. *G. W. McConkie and K. Rayner, "The span of the effective stimulus during a fixation in reading,"* Perception & Psychophysics **17**, *pp. 578–586, 1975.*
17. *L. C. Loschky and G. W. McConkie, "User performance with gaze contingent multiresolutional displays," in* Proceedings of Eye Tracking Research & Applications, *pp. 97–103, 2000.*
18. *P. Kortum and W. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," in* Human Vision and Electronic Imaging, SPIE Proceedings, **2657**, *pp. 350–360, 1996.*
19. *W. S. Geisler and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," in* Human Vision and Electronic Imaging: SPIE Proceedings, *B. Rogowitz and T. Pappas, eds., pp. 294–305, 1998.*
20. *J. S. Perry and W. S. Geisler, "Gaze-contingent real-time simulation of arbitrary visual fields," in* Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA, *B. E. Rogowitz and T. N. Pappas, eds.,* **4662**, *pp. 57–69, 2002.*
21. *M. Dorr, M. Böhme, T. Martinetz, and E. Barth, "Visibility of temporal blur on a gaze-contingent display," in* APGV 2005 ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization, *pp. 33–36, 2005.*