Guiding Eye Movements for Better Communication and Augmented Vision

Erhardt Barth[1], Michael Dorr[1], Martin Böhme[1], Karl Gegenfurtner[2], and Thomas Martinetz[1]

[1] Institute for Neuro- and Bioinformatics, University of Lübeck,
Ratzeburger Allee 160, D-23538 Lübeck, Germany
{barth, dorr, boehme, martinetz}@inb.uni-luebeck.de
http://www.inb.uni-luebeck.de
[2] Allgemeine Psychologie, Justus-Liebig-University,
Otto-Behaghel-Str. 10, D-35394 Gießen, Germany
Karl.R.Gegenfurtner@psychol.uni-giessen.de

Abstract. This paper briefly summarises our results on gaze guidance such as to complement the demonstrations that we plan to present at the workshop. Our goal is to integrate gaze into visual communication systems by measuring and guiding eye movements. Our strategy is to predict a set of about ten salient locations and 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, in our current implementation, we show a natural-scene movie and overlay red dots very briefly such that they are hardly perceived consciously. To decrease the probability, for example, we locally reduce the temporal frequency content of the movie. We here present preliminary results, which show that the three steps of our above strategy are feasible. The long-term goal is to find the optimal real-time video transformation that minimises the difference between the actual and the desired eye movements without being obtrusive. Applications are in the area of vision-based communication, augmented vision, and learning.

1 Introduction

An important property of human vision is that we must constantly shift our gaze between objects of interest because only the central part of the retina provides high visual acuity. Moreover, 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 message 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.

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 human visual capabilities. Such augmented-vision systems are of particular interest for automotive 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. During the workshop we will demonstrate the eye-tracking and display technology that we currently use and continue to develop [1, 2]. In the remainder of this paper we report on a few results that address selected problems of gaze guidance.

2 Guidance of Eye Movements

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 for saccade targets, (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 the individual points (i-iii) above.

2.1 Eye Movement Predictions

We have investigated the variability of eye movements with dynamic natural scenes and found that 5-15 clusters (frequently looked at regions that, when taken together, cover 2-5 % of the viewing area) account for 60 % of all fixations [3]. This justifies our approach of predicting a small set of candidate locations based on low-level features of the visual input.

Like other authors, e.g. [4], 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. We assume 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 (although in reality, bottom-up and top-down processing will most likely be hard to segregate). 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 [5, 6], our approach to saliency is based on the concept of intrinsic dimension that was introduced for still images in [7]. 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 [8]. 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 [10]. The structure tensor is defined in terms of the spatio-temporal gradient ∇f of the image intensity function

$$\mathbf{J} = \omega * \nabla\mathbf{f} \otimes \nabla\mathbf{f} = \omega * \left(\nabla\mathbf{f}\nabla\mathbf{f}^{\mathbf{T}}\right) \tag{1}$$

where ω denotes a convolution kernel that performs a local averaging of the product terms. Our saliency measure that was used for the results presented in Fig. 1 is K, the determinant of $\mathbf{J}$, which indicates an intrinsic dimension of 3, i.e., there is no spatio-temporal direction along which intensity is constant. Typical features with high K-values would be blobs or corners that appear or dissappear. 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. Details of how we select the candidate locations from the saliency map are given in [11]. We have used other invariants of the structure tensor (trace and sum of minors that indicate intrinsic dimensions of at least one or two respectively) and found prediction results that were slightly worse. We have not performed more comprehensive comparisons for two reasons: (i) the results are good enough for our purpose of guiding eye movements, and (ii) we are not aware of similar attempts to predict a set of locations based on a spatio-temporal measure.

As a baseline for assessing the saliency maps computed analytically using the K measure, we use empirical saliency maps, i.e. saliency maps computed from the actual eye movements of the test subjects. These 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 actually does even better than the best possible saliency because it is 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 [6].

Note that the attempt of predicting ten candidate locations based on a simple saliency measure is quite successful in the sense that it approaches the limit given by the empirical predictor. However, when using only one instead of ten candidates, the errors obtained for both the empirical and the analytical saliency measures are still better than chance but unacceptably high.

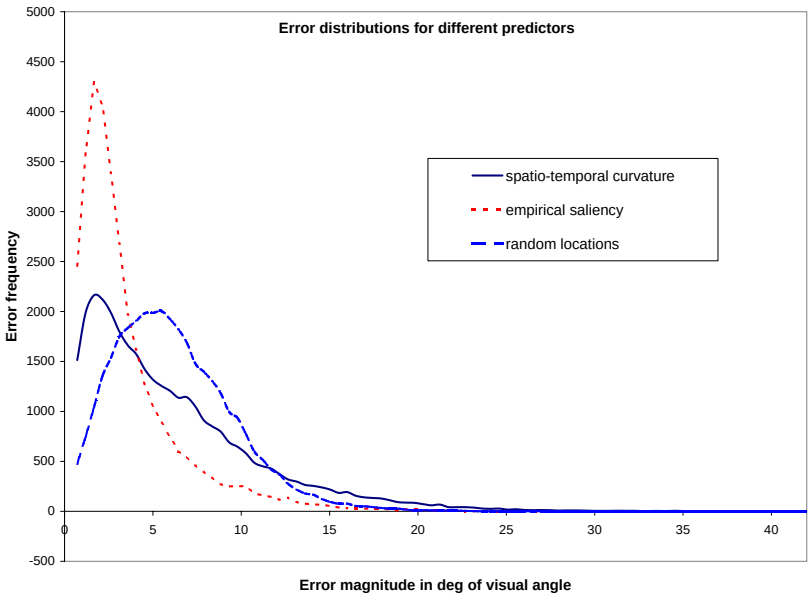

Fig. 1. Saliency prediction results. Histogram of error (distance of saccade target to closest salient candidate location) for ten candidate locations. The horizontal axis plots the error magnitude in degrees, the vertical axis plots the number of saccades per histogram bin. Plots are shown for the K saliency measure, the empirical saliency measure, and locations chosen at random.

2.2 Effect of Gaze-Contingent Red Dots

We now address the problem of increasing the probability of candidate locations to be attended. In a first set of experiments, we were motivated by the well-known fact that sudden object onsets in the visual periphery can attract attention. We therefore chose to briefly superimpose small bright red dots on the displayed 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. 2. Note that the red-dot stimuli have a considerable effect when presented at 10 degrees eccentricity. The effect is smaller at 15 and 20 degrees partly because of the limited field of view and partly due to the fact that the stimulus size was not scaled (enlarged by the cortical magnification factor) with eccentricity and was therefore less effective.

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. In another set of experiments, where (other) subjects watched the same movies and were instructed to look for red dots and press a button when they detected one, the stimulation remained invisible in about 50 % of cases. Stimuli at 15 and 20 degrees were less visible than those at 10 degrees.

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.

2.3 Effect of Gaze-Contingent Spatio-Temporal Filtering

We now address the problem of decreasing the probability of candidate locations to be attended. To do this, 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 [12].

Gaze-contingent displays manipulate some property of the (static or moving) image as a function of gaze direction (see [13] for a review). 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 [14, 15]. The current state-of-the-art algorithm for gaze-contingent spatial filtering of video is due to Perry and Geisler [16]. Unlike previous algorithms, which introduced artifacts in the filtered images, their algorithm produces smooth, artifact-free results.

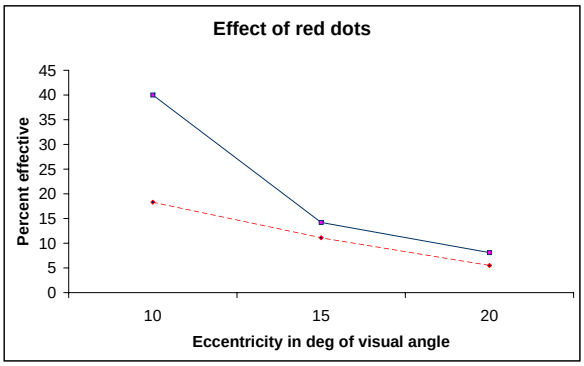

Fig. 2. Effect of red-dot stimulation. We show the number of cases (in % of all trials) in which subjects made a saccade to where the red dot had been shown (the saccade started in a time window of 100 to 300 ms after stimulus onset and landed within a circle of 5 deg radius around stimulus location). As a reference we show (with a dashed line) the same results for cases in which the red dots had not been displayed.

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 (demonstrations will be shown at the workshop). 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. 3 show that gaze-contingent temporal filtering reduces the number of saccades into the periphery where the temporal frequencies are reduced by our gaze contingent display. 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.

3 Discussion

We have shown that a rather small set of locations 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 previous analysis of eye movements on high resolution natural videos, which shows that eye movements tend to cluster in rather few (10-20) locations [3]. We then reported on simple experiments that were meant to increase the saliency by a brief gaze-contingent presentation of red dots in the periphery. Concerning methods for

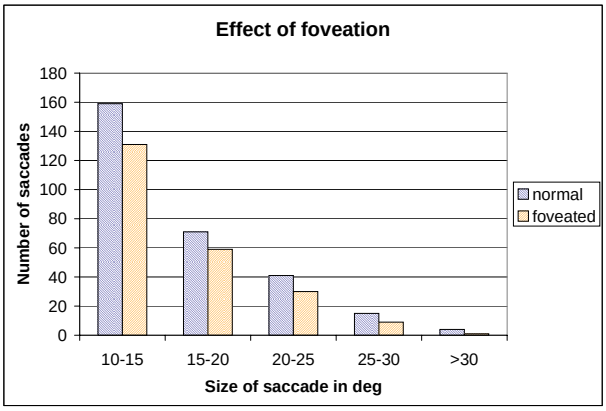

Fig. 3. Effect of spatio-temporal filtering shown as histograms of saccade amplitudes with and without gaze-contingent temporal filtering. Only the histogram bars for saccades of 10 degrees and greater are shown since no effect could be expected in the central field of view due to the shape of the foveation function.

decreasing saliency, we have shown that peripheral 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 red dots and of the temporal blur. The visibility is hard to determine since it depends on the task. However, our results [17] indicate that there exists a set of manipulations (of which we do not yet know how large it is) that are effective in guiding eye movements but are not perceived consciously.

We therefore conclude that gaze guidance seems possible in principle. However, 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 thus justify the applications that we have in mind.

Acknowledgements

Research was supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01B with acronym *ModKog. We thank the reviewers for valuable remarks.*

References

1. *Dorr, M., Böhme, M., Martinetz, T., Barth, E.: Gaze-contingent spatio-temporal filtering in a head-mounted display. In: Perception and Interactive Technologies. (same volume)*

2. *Meyer, A., Böhme, M., Martinetz, T., Barth, E.: A single-camera remote eye tracker. (Same volume)*
3. *Dorr, M., Böhme, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Analysing and reducing the variability of gaze patterns on natural videos. In Groner, M., Groner, R., Müri, R., Koga, K., Raess, S., Sury, P., eds.: Proceedings of 13th European Conference on Eye Movements. (2005) 35*
4. *Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259*
5. *Barth, E., Drewes, J., Martinetz, T.: Individual predictions of eye-movements with dynamic scenes. In Rogowitz, B., Pappas, T., eds.: Electronic Imaging 2003. Volume 5007., SPIE (2003) 252–259*
6. *Böhme, M., Dorr, M., Krause, C., Martinetz, T., Barth, E.: Eye movement predictions on natural videos. Neurocomputing (2006) (in press).*
7. *Zetzsche, C., Barth, E.: Fundamental limits of linear filters in the visual processing of two-dimensional signals. Vision Research **30** (1990) 1111–1117*
8. *Mota, C., Barth, E.: On the uniqueness of curvature features. In Baratoff, G., Neumann, H., eds.: Dynamische Perzeption. Volume 9 of Proceedings in Artificial Intelligence., Köln, Infix Verlag (2000) 175–178*
9. *Barth, E., Watson, A.B.: A geometric framework for nonlinear visual coding. Optics Express **7** (2000) 155–165*
10. *Jaehne, B., Haußecker, H., Geißler, P., eds.: Handbook of Computer Vision and Applications. Academic Press (1999)*
11. *Böhme, M., Dorr, M., Krause, C., Martinetz, T., Barth, E.: Eye movement predictions on natural videos. Neurocomputing (2005) (in press).*
12. *Böhme, M., Dorr, M., Martinetz, T., Barth, E.: Gaze-contingent temporal filtering of video. In: Eye Tracking Research and Applications (ETRA). (2006) (in press).*
13. *Duchowski, A.T., Cournia, N., Murphy, H.: Gaze-contingent displays: A review. CyberPsychology & Behavior **7** (2004) 621–634*
14. *Kortum, P., Geisler, W.: Implementation of a foveated image coding system for image bandwidth reduction. In: Human Vision and Electronic Imaging, SPIE Proceedings. Volume 2657. (1996) 350–360*
15. *Geisler, W.S., Perry, J.S.: A real-time foveated multiresolution system for low-bandwidth video communication. In Rogowitz, B., Pappas, T., eds.: Human Vision and Electronic Imaging: SPIE Proceedings. (1998) 294–305*
16. *Perry, J.S., Geisler, W.S.: Gaze-contingent real-time simulation of arbitrary visual fields. In Rogowitz, B.E., Pappas, T.N., eds.: Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA. Volume 4662. (2002) 57–69*
17. *Dorr, M., Böhme, M., Martinetz, T., Barth, E.: Visibility of temporal blur on a gaze-contingent display. In: APGV 2005 ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization. (2005) 33–36*