Predicting, analysing, and guiding eye movements

Michael Dorr[*]
Institute for Neuro- and Bioinformatics
University of Lübeck
Germany
dorr@inb.uni-luebeck.de

Martin Böhme
Institute for Neuro- and Bioinformatics
University of Lübeck
Germany
boehme@inb.uni-luebeck.de

Thomas Martinetz
Institute for Neuro- and Bioinformatics
University of Lübeck
Germany
martinetz@inb.uni-luebeck.de

Erhardt Barth
Institute for Neuro- and Bioinformatics
University of Lübeck
Germany
barth@inb.uni-luebeck.de

Abstract

In this paper, we will present an overview of our work that is aimed at integrating gaze into visual communication systems by measuring and guiding eye movements [1]. This requires investigating and modelling how eye movements are determined by the visual input, modelling what is relevant to the user, and new technological developments for better eye tracking and fast gaze-contingent graphics. A number of challenges remain, some of which may be solved by machine-learning techniques, e.g. predicting eye movements and inferring a person's intent.

1 Background

Vision is the dominant perceptual channel through which we interact with information and communication systems, but one major limitation of our visual communication capabilities is that we can attend to only a very limited number of features and events at any one time (e.g., [2]). This fact has severe consequences for visual communication, because what is effectively communicated depends to a large degree on those mechanisms in the brain that deploy our attentional resources and determine where we direct our eye movements, i.e. our gaze.

Therefore, future information and communication systems should use gaze guidance to help the users deploy their limited attentional resources more effectively. Gaze guidance here means that the user follows a prescribed pattern with their gaze, thus taking in information in a specific, potentially more efficient way.

When dealing with the problem of guiding a user's gaze, we face several challenges. In this paper, we will give an overview of the issues that we have begun to address so far, but there remain a number of open problems.

[*] http://www.inb.uni-luebeck.de

The first challenge is that we need to analyse in more detail how humans watch dynamic scenes. The majority of previous research on eye movements has dealt with static scenes only, mainly because of the technical problems inherent in recording eye movements on movies. However, we believe that it is more practicable to determine what drives eye movements in dynamic scenes. Bottom-up features, that is features that are directly computable from an image sequence such as brightness, colour, or motion, should have a greater influence on directing gaze here than in static scenes. Accordingly, attempts have recently been made at modelling what low-level features determine eye movements in moving scenes as well [3]. Based on these findings, we have been able, at least to some extent, to predict where observers will direct their gaze from a number of previously attended locations [4].

A further requirement is a quality function that estimates how well-suited an observer's gaze pattern is for a given image sequence. There are basically two approaches to obtain an optimal gaze pattern. It is well known that experts, for example experienced car drivers, employ viewing strategies different from those of novices. Thus, the gaze pattern of an expert could be recorded and "replayed" to the user. The more generic approach uses an image-processing algorithm that could identify the most informative regions in a scene. Of course, we also need a model of both the task at hand and the observer's intentions to decide which information might be relevant to the observer.

Finally, the eye movements actually need to be guided to follow the intended gaze pattern. In a strict sense, this will be impossible to achieve, as an observer might consciously choose to only focus on one specific aspect of a scene. Nevertheless, we believe that for most purposes, it will suffice to significantly increase the likelihood that certain locations will be fixated, while suppressing other potential saccade targets. Indeed, we have developed a number of spatio-temporal transformations that, as we were able to show, change the eye movements of observers, although the guidance still needs to become more specific.

When all these challenges have been met, we furthermore not only want to change the observer's eye movements, but also achieve a change in behaviour, i.e. an improvement in actual task performance. In the next section, we will give an overview of potential applications.

2 Applications

An important potential application of gaze-guidance systems is augmented vision. Augmented-vision systems can be designed to integrate human vision and computer vision. For example, in a car, the driver's attention can be directed towards a pedestrian who has been detected by sensors looking out of the car.

A further application is the use in training systems. 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 recording the gaze pattern of experts and applying it to novices, we can evoke a sub-conscious learning effect.

Finally, current technical visual communication systems are based on the physical properties of images and cannot improve the communication process as such, because they do not address the question of what message is conveyed by an image or a video. Future communication systems' images and movies can be defined not only by brightness and colour, but will also be augmented with a recommendation of where to look, of how to view the images. For this to succeed, such systems will also have to take into account the user's intentions.

3 The machine-learning perspective

The human visual system is highly complex. This complexity is increased even further when we extend our view to include the higher cognitive functions that also play a role in controlling the direction of gaze, such as alertness, emotional state, or intent. Therefore, we believe that it will be impossible to distill a set of fixed parameters that will allow gaze guidance to work in all situations, for all users. Rather, we believe that the gaze-guidance display will have to continuously adapt to the user and the task at hand.

In the gaze-guidance display, what is displayed is a function of the user's gaze, while at the same time the display influences the user's gaze. In this closed loop, there exists a multitude of parameters that need to be adjusted in an on-line fashion. Different users may have different physiological characteristics, such as saccadic latency, or different cognitive strategies, attentional states, or expectations. Lastly, the search space of spatio-temporal transformations that might possibly be used to guide gaze is too vast to be explored systematically. We have implemented some basic transformations (see section 4), but fine-tuning will have to be done in an unsupervised, continuously evolving manner.

The following sections will outline some of the issues we have addressed so far with machine-learning techniques and other methods.

3.1 Analysis of eye movements

We have investigated the variability of eye movements on dynamic natural scenes [5]. To this end, we collected a large data set of gaze samples from 54 subjects watching a variety of short 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 centered 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" which contain a high number of fixation locations. On average, 5-15 clusters (2-5% of the viewing area) account for 60% of all fixations (see Fig. 1 for an example).

3.2 Eye movement predictions

The model we use to predict where an observer is going to look is composed of two distinct parts. This separation is motivated by the two fundamental types of eye movements that are relevant to our purposes. First, saccades are ballistic high-velocity eye movements that serve to move the fovea from one fixation location to another; during a saccade, most of the visual input is suppressed so that, for example, we do not perceive the blur induced by the motion of the visual scenery across the retina. We define the task of saccade prediction as predicting the target of the saccade, not the complete saccade trajectory, because the latter is irrelevant for our purposes. The second type of eye movements comprises all movements that are made between saccades. These movements can be further classified into a variety of types, but for the present purpose, they share the common characteristic that velocity is relatively low. To model such intersaccadic eye movements, we use a supervised-learning technique that, from a history of previous gaze samples, predicts the gaze position in the next time step.

For the prediction of saccade targets, one ideally would be able to predict a single location that has a high probability of being the next saccade target. To achieve this, however, a complete understanding of the higher decision processes involved in saccadic programming would be required. For example, an observer might choose to fixate one part of a scene over another for purely semantic reasons. Therefore, we restrict ourselves to predicting only a certain number of locations that are likely to be fixated as the next saccade target. We have

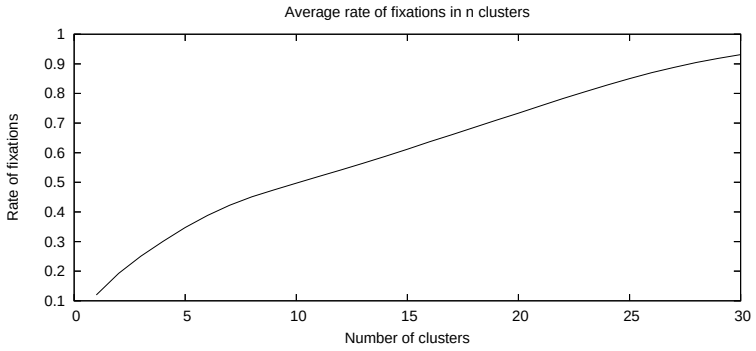

Figure 1: Rate of fixations that fall into the first n extracted clusters, for one example movie. These rates were computed on a per-frame basis and averaged across the whole movie.

attempted to use machine learning techniques to predict which of the candidate locations will be chosen, but with only limited results so far.

To extract candidate locations for saccade targets from a video, we use a saliency map that assigns a certain degree of saliency to every location in every frame of the video sequence. Various techniques exist for computing saliency maps, most of which are intended to model the processes in the human visual system that generate potential saccade targets [3].

The saliency map used here is based on the spatio-temporal curvature of the image sequence. The curvature is computed here using the determinant of the structure tensor, which is defined as the locally averaged outer product of the (x, y, t) intensity gradient. To avoid all candidate locations being extracted from only a single small high-curvature region in an image, we extract locations by iteratively applying a lateral inhibition algorithm, so that locations with a high saliency close to a local maximum become suppressed.

Fig. 2 compares the performance of our predictor with that of a predictor that uses an empirical saliency map, which is derived from the recorded eye movement data as described in section 3.1: Clusters with a high density of fixations are assigned a high saliency. This empirical saliency map gives an upper bound of what we can expect to achieve with a purely bottom-up approach, without modelling the user's top-down influences.

The results show that, currently, the performance of our predictor is about halfway between the results one would obtain by guessing locations at random and that of the ideal predictor based on the empirical saliency map. For a detailed discussion of our predictor, see [4].

4 Current state of the art

The final goal of our gaze guidance system is to direct the user's attention to a specific part of a scene without the user noticing this guidance. Apart from our work on modelling which image features attract gaze, we have therefore also conducted experiments with sev-

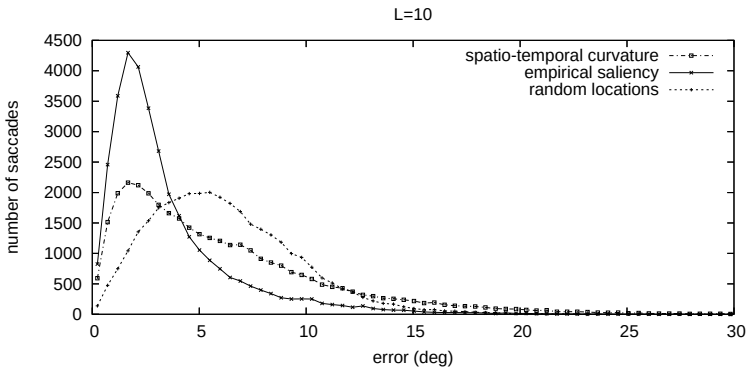

Figure 2: Error distributions for different predictors. The error is the distance of the recorded saccade target to the closest of $L = 10$ predicted candidate locations.

eral 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. In about 50% of trials, saccades were initiated towards the location of the flashed red dot. 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 65% of cases, this stimulation remained invisible. Similar results were obtained in an experiment where the red dot was replaced by a looming stimulus. 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, ideally by an automated learning process.

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 [6], based on earlier work that manipulated only spatial resolution [7]. For example, we can selectively blur high temporal frequencies in the visual periphery, which are known to evoke saccades. Because of the limited perceptual capabilities of the human visual system in the periphery, this blur remains unnoticed. Nonetheless, we were able to show that such peripheral temporal blur suppresses saccades towards the periphery. Next, we plan to specifically change the spatio-temporal content only at certain locations in an image.

5 Conclusion

We have here described the efforts we have made to not only infer and predict human behaviour (eye movements) but also change it such as to improve human performance. Preliminary results indicate that it should be possible to guide the gaze of a person [8]. A number of problems that need to be solved can be addressed by machine-learning techniques. The ultimate goal would be to find the optimal way to display information such as to minimize the error between the actual and the desired performance of a person performing certain actions in a particular environment, e.g. to avoid traffic accidents, or the difference between the information that is intended to be received and the one that is actually received, e.g. by a person watching a movie or a news programme.

Acknowledgments

Research is supported by the German Ministry of Education and Research (BMBF) under grant number 01IBC01, ModKog.

References

[1] Information technology for active perception homepage, 2001. http://www.inb.uni-luebeck.de/Itap/.

[2] J Kevin O'Regan, Ronald A Rensink, and James J Clark. Change-blindness as a result of 'mudsplashes'. *Nature*, 398:34, 1999.

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

[4] Martin Böhme, Michael Dorr, Christopher Krause, Thomas Martinetz, and Erhardt Barth. Eye Movement Predictions on Natural Videos. *Neurocomputing*, 2005. (in press).

[5] Michael Dorr, Martin Böhme, Jan Drewes, Karl R Gegenfurtner, and Erhardt Barth. Variability of eye movements on high-resolution natural videos. In Heinrich H Bülthoff, Hanspeter A Mallot, Rolf Ulrich, and Felix A Wichmann, editors, *Proceedings of the 8th Tübinger Perception Conference*, page 162, 2005.

[6] Michael Dorr, Martin Böhme, Thomas Martinetz, and Erhardt Barth. A gaze-contingent display with variable temporal resolution. In *Proceedings of the BIP Workshop on Bioinspired Information Processing, Lübeck, Germany*, page 18, 2005.

[7] Jeffrey S Perry and Wilson S Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In B E Rogowitz and T N Pappas, editors, *Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA*, volume 4662, pages 57–69, 2002.

[8] Michael Dorr, Thomas Martinetz, Karl Gegenfurtner, and Erhardt Barth. Effects of gaze-contingent stimulation on eye movements with natural videos. *Perception Suppl.*, 33:134, 2004.