Eye movement modelling and gaze guidance

Michael Dorr[1], Eleonora Vig[1], Karl R. Gegenfurtner[2], Thomas Martinetz[1], and Erhardt Barth[1]

Abstract We show that a simple low-dimensional representation of movie patches, namely local spectral energy, can be used to predict where people will look in dynamic natural scenes. We then present a gaze-contingent display that modifies local spectral energy in real time. This modification of the saliency distribution of the scene leads to a change in eye movement statistics. Our research aims at the guidance of gaze with the ultimate goal of optimising vision-based communication systems.

1 INTRODUCTION

The eye movements made by observers on a visual scene are tightly linked to their perception. A well-studied corollary of this fact is that experts employ different gaze patterns to solve visual tasks than novices, e.g. in driving a car, controlling air traffic, or scanning X-rays [6, 10].

However, gaze direction and gaze patterns are typically not taken into account in today's information and communication systems. With the advent of ever cheaper and more robust eye-tracking technology, we propose that gaze patterns should become an integral image attribute similar to the physical image attributes luminance and colour. To this end, gaze patterns need to be sensed and displayed. For the former, eye trackers are already commercially available; for the latter, we propose gaze-contingent interactive displays that modify visual content in real time to guide the observer's gaze. Such guidance would ultimately allow novices to "see with the eyes of experts" and to optimise vision-based communication in general. Of particular interest for human-computer conversation is the possibility to combine gaze guidance with further modalities such as speech and emotions.

In this paper, we will outline the strategy that we envisage to implement gaze guidance and some initial results that were obtained with a prototype gaze-contingent interactive display. We will start with a description of our work on modelling and predicting eye movements and attention. The question of what image features draw fixations has a long tradition in vision science. A common approach is to derive biologically inspired feature extraction algorithms, such as contrast or edge detectors, and to compute several such features for each location of an image or image sequence. From a combination of these feature maps, a so-called saliency map finally can be formed that indicates how salient or interesting any location is, and those regions where saliency values are high are used to predict where fixations will occur [3, 8]. Only recently, attempts have been made to use Machine Learning algorithms to automatically extract relevant features from large sets of eye movement data [9]. Following this approach, we trained a support vector machine with the spectral energy distribution of fixated and non-fixated image regions and achieved up to 79% prediction accuracy (AUC score), a very favourable result compared to other published prediction algorithms.

We then used these results to design a gaze-contingent interactive display that locally modifies spectral energy to change the saliency distribution across the visual scene. To guide the observer's gaze to a specific location, the saliency at this desired location should be increased, whereas the saliency of possible distractors everywhere else should be reduced. The difficult problem here is to find the psychophysically optimal transformation that makes image regions more or less salient: reducing luminance, for example, might intuitively seem to be an effective reduction in saliency, but a dark spot in an otherwise normally-lit scene could attract attention in itself. Any low-level modification of natural scenes also needs to take into account the expectations of an observer: unnatural-looking regions might be both salient and disturbing. For example, it has been shown that gaze guidance is possible with flashing red dots [1], but such manipulation cannot be embedded well in natural movies. Besides these perceptual issues, the technical implementation of such image processing in real time and as a function of gaze direction also is a major challenge. In a first experiment with our prototype gaze-contingent display, we were indeed able to alter eye movement statistics; however, the intended effect of gaze guidance suffered from technical artefacts that will be discussed.

2 MODELLING OF EYE MOVEMENTS AND ATTENTION

We used an SR Research EyeLink II eye tracker running at 250 Hz to record eye movement data from 54 subjects watching 18 high-resolution videos of natural outdoor scenes. Trials where more than 5% of gaze samples were invalid (typically, because the subject blinked excessively) were discarded, leaving 844 trials for further analysis. Videos had a duration of about 20 s each, a spatial resolution of 1280 x 720 pixels, and were shown at 29.97 frames per second on a screen covering 48 x 27 degrees of visual angle, so that the maximum displayed frequency was 13.4 cycles per degree. Overall, about 40,000 saccades were extracted from the raw data using a velocity-based two-step procedure. The landing points of these saccades and their spatio-temporal neighbourhoods represented the class of fixated regions; for the class of non-fixated regions that were of relatively low saliency, we shuffled the movies and their corresponding gaze data, so that the non-fixated regions in movie A were picked from the saccade landing points in movie B and vice versa. This standard approach might lead to some overlap in the classes of fixated and non-fixated regions, but factors out any bias in the spatial distribution of objects in our movies and the tendency of human subjects to look at the centre of the screen rather than the periphery.

For each image region in the two classes, local spectral energy was computed based on a spatio-temporal Laplacian pyramid.

[1] Inst. for Neuro- and Bioinformatics, Univ. of Lübeck, GER.
Email:{dorr, vig, martinetz, barth}@inb.uni-luebeck.de

[2] Abt. für Allgemeine Psychologie, Justus-Liebig-Univ. Gießen, GER.
Email: karl.r.gegenfurtner@psychol.uni-giessen.de

The Laplacian pyramid [4] is an efficient bandpass representation for image sequences and can be computed from the Gaussian pyramid, which contains the original sequence and successively smaller versions of the input. These smaller versions are created by reducing resolution in each spatio-temporal dimension by a factor of two (i.e. every other pixel or frame is thrown away). Because the maximum frequency that can be represented in an image corresponds to its resolution, a lowpass filter has to be applied before this downsampling step; as a consequence, the smaller versions of an image sequence also contain only (the low-frequency) parts of its frequency spectrum (see Fig. 1, left). Adjacent levels of the Gaussian pyramid are then brought back to the same resolution (by interpolating every other pixel and frame of the lower-resolution level) and subtracted from each other, so that the Laplacian pyramid finally consists of single bandpass frequency bands (see Fig. 1, right). Because the lower frequency bands are not stored at the full resolution of the original image sequence, this representation can efficiently be utilised even for a high number of pyramid levels (a spatial pyramid in the limit has only 1+1/4+1/16+...= ~1.33 times as many pixels as the original image; because a temporal pyramid is downsampled in only one dimension, this factor is 1+1/2+1/4+...= ~2).

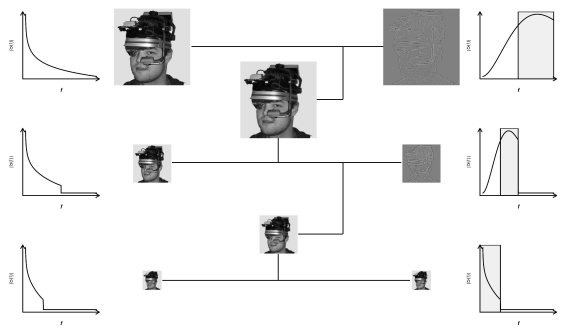

Figure 1. The Laplacian pyramid (here, spatial dimension only): first, a Gaussian pyramid is created by successively filtering and downsampling the original image (images left). The lower-resolution versions only contain the low-frequency parts of the spectrum (graphs left). Subtracting adjacent levels yields single frequency bands (right).

To obtain local spectral energy, the average squared pixel intensity in the neighbourhood of each saccade landing point was computed on each level of a Laplacian pyramid with 5 spatial and 5 temporal levels, so that a single training sample was a 25-dimensional vector. Theoretically, we could have used those single pixels that corresponded to the saccade landing point directly; but because of the spatio-temporal imprecision and noise inherent in both the human visual system and the eye tracking equipment, we averaged over a window with a size of about 5 degrees of visual angle. This number was found by variation of this parameter and is in line with previous studies on the optimal window size for prediction [9]. Although averaging energy instead of using all pixels in the window for a training sample reduces the information available to the predictor, the simultaneous reduction in dimensionality actually improves classification results. Then, a soft-margin support vector machine was trained with two thirds of all available samples and the optimal parameters were found with cross-validation (using the publicly available libsvm package [5]). Generalisation performance was then tested on the remaining third of samples; we achieved a very favourable AUC (area under the curve) score of about 0.79.

3 GAZE-CONTINGENT INTERACTIVE DISPLAY

Gaze-contingent displays change their content as a function of gaze direction. The first such displays used very simple modifications; for example, a moving mask that would display text only around the centre of fixation and completely blank out the periphery was used to determine the perceptual span in reading [11]. Perry and Geisler introduced an algorithm based on a spatial Gaussian multiresolution pyramid that can simulate arbitrary visual fields [12]. The Gaussian pyramid of an image sequence is computed in real time; each level stores the images with a different resolution and, therefore, a different maximum frequency. By interpolating between adjacent levels of the pyramid during generation of the output image, the maximum spatial resolution can be specified for each pixel in retinal coordinates. Böhme et al. developed a similar display that operates in the temporal domain [2], so that moving objects in the periphery are almost erased, whereas the stationary background stays intact. Indeed, this manipulation is not visible to an observer, but reduces the number of large saccades (into the periphery) [1]. We now extended this algorithm to be based on a Laplacian pyramid, which gives us access to single frequency bands; instead of only specifying a cutoff frequency per pixel, local multiplications can freely alter the frequency response of the system and thus change local spectral energy.

Because of computational constraints, we used a spatial Laplacian pyramid only. This has two severe disadvantages. First, the temporal dimension is particularly important in determining what image regions are salient (e.g. peripheral motion is a strong attractor for attention). Second, a multiscale approach inherently allows for smooth transitions between modified and unmodified locations; using just one temporal level means that the onset of a modification happens abruptly from one frame to the next.

To assess the effect our gaze-contingent display had on eye movement statistics, 12 subjects watched 6 movies (out of the 18 movies used for prediction, see above; they were slightly downsampled to a resolution of 1024 x 576 pixels to meet real-time demands) each. From the eye movement recordings on unmodified videos (see above), we determined the 20 "candidate" locations per frame most likely to be fixated in the near future (we set this anticipated latency to 100 ms, roughly the minimal time required for programming a saccade). In principle, our prediction algorithm presented above could be used to find these candidate locations as well; however, for assessing only the gaze-guidance effect of our display, we decided to use the best predictor available (namely, where other people looked). During each trial, the movie was decomposed into a spatial Laplacian pyramid with 5 levels in real time. After each saccade a subject made, local spectral energy was reduced at each candidate location that fell into three randomly chosen quadrants of the subject's visual field (the remaining quadrant was left unmodified; so were the central 5 degrees around

fixation to prevent the subject from consciously perceiving any manipulation). On the first four pyramid levels, energy was reduced to 1.2 times the average energy of non-fixated locations, which corresponds to a mean reduction by a factor of 1.6 (see Fig. 5) -- the lowest level or DC component, which contains the mean luminance, was not changed to avoid visible artefacts. A window of 9 x 9 pixels was modified, corresponding to 3.4 x 3.4 degrees on the lowest modified level. Then, the pyramid was synthesized in less than 10 ms and the resulting frame was displayed on the screen (for an example, see Fig. 2-4). Including the latency of the eye tracker and that of the computer screen, the overall latency was about 20-30 ms from the end of a saccade to the resulting change on the display. Because of this short latency, saccadic suppression (the suppression of visual input around the time of a saccade, while the whole visual world is moving across the retina at high speed) should reduce the visibility of these changes to the subjects or even render them unnoticeable.

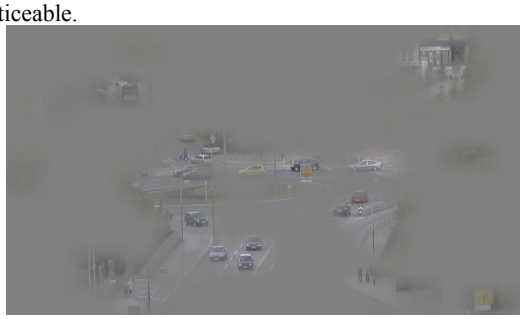

Figure 2. Example fixation map created from eye movement recordings of 54 people. Non-gray regions indicate where people looked; this data was used to predict candidate locations.

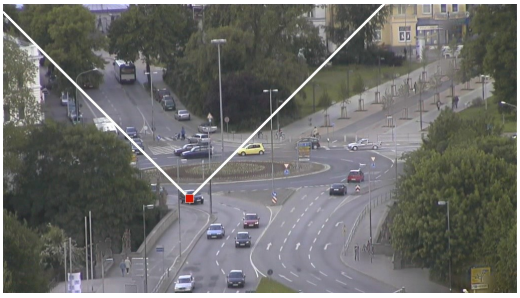

Figure 3. Example stillshot from GCID experiment. Gaze position at intersection of lines; lines mark unmodified quadrant (markers not shown during actual trials). Note the reduced contrast at (e.g.) the road signs bottom right or in the centre of the roundabout -- cf. Fig. 4.

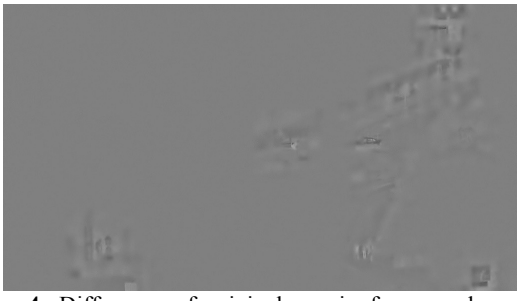

Figure 4. Difference of original movie frame and modified version as shown on the gaze-contingent display. Note the correspondence with the fixation map (see Fig. 2).

4 RESULTS

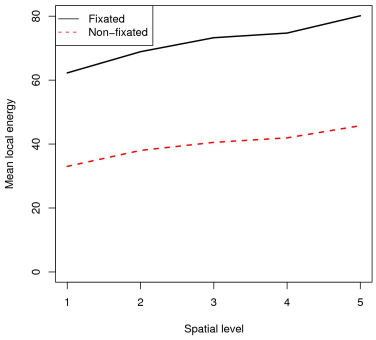

Figure 5. Local spectral energy. Average energy is almost twice as high for attended as for non-attended locations.

As a baseline measure, we performed the same experiment as above, but instead of using 'live' data from real subjects, we replayed the eye movement recordings on unmodified videos. Therefore, stimulation placement, timing, etc. were identical to our experiment data, but the stimulation could not possibly have had an effect. Ideally, the reduction of local spectral energy in the three modified quadrants should have made these quadrants less salient, leading to proportionally more fixations in the unmodified quadrant. But this was not the case (see Fig. 6); however, subjects reported having seen an occasional flicker on the screen, and our data indicates that it was this flicker -- when the graphics update was too slow -- that actually had the opposite effect and attracted fixations. Nevertheless, one effect of our display was a reduction of overall saccade rate of up to 20% (see Fig. 7); a finding reflected by an increase in fixation durations (see Fig. 8).

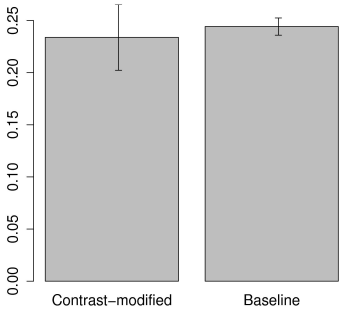

Figure 6. Rate of saccades into unmodified quadrant. In the baseline condition, about 25% of all saccades fall into this quadrant; in trials with our gaze-contingent display, this rate is slightly lower (although the increase in saliency relative to the remaining 3 quadrants should have increased this rate) and shows a higher variance.

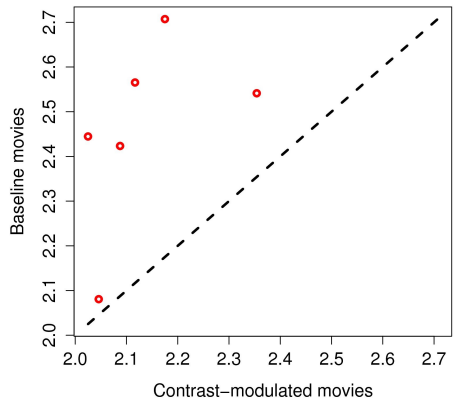

Figure 7. The number of saccades per second decreased significantly when local spectral energy was reduced at highly salient points (Wilcoxon test, $p<0.032$). Each point represents one of the six movies shown on the gaze-contingent display.

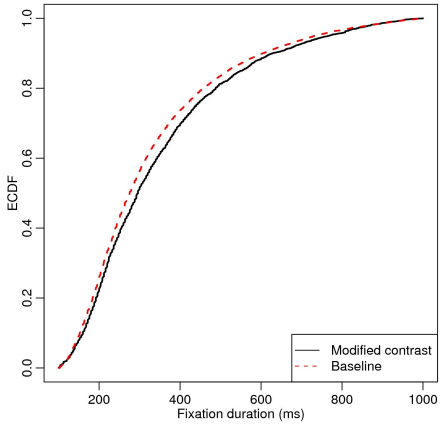

Figure 8. Empirical cumulative distribution function of fixation durations. Fixations are significantly longer on the gaze-contingent display (Kolmogorov-Smirnov test, $p<<0.01$).

5 CONCLUSIONS AND OUTLOOK

We have shown that a low-dimensional representation of movie patches, namely local spectral energy, can be used successfully to predict where humans will fixate in a natural dynamic scene. This result is useful not only in modelling the human visual system, but should also have broader implications in the design of human-machine interaction. We furthermore showed that we are now able to perform fairly sophisticated modifications of dynamic scenes in real time and that such modifications can alter eye movement statistics. Currently, we are working towards even faster image processing routines implemented in commodity graphics hardware [7]. This will allow us to achieve lower latencies and, therefore, simultaneously reduce visibility of the modifications and create some headroom for more complex transformations. We are also extending our approach into the temporal dimension; this is computationally more complex, but also promises to be much more effective perceptually, to ultimately enable a gaze-guidance effect.

REFERENCES

[1] E. Barth, M. Dorr, M. Böhme, K. R. Gegenfurtner, and T. Martinetz. Guiding the mind's eye: improving communication and vision by external control of the scanpath. In: *Human Vision and Electronic Imaging*, volume 6057 of *Proc. SPIE*. B. E. Rogowitz, T. N. Pappas, and S. J. Daly (Eds.) (2006).

[2] M. Böhme, M. Dorr, T. Martinetz, and E. Barth. Gaze-contingent temporal filtering of video. In *Proceedings of Eye Tracking Research & Applications (ETRA)*, 109-115 (2006).

[3] N. D. B. Bruce, D. P. Loach, and J. K. Tsotsos. Visual correlates of fixation selection: a look at the spatial frequency domain. In: *ICIP 2007*, 289–92 (2007).

[4] P. J. Burt and E. H. Adelson, The Laplacian Pyramid as a Compact Image Code. *IEEE Transactions on Communications*, 31(4):532-540 (1983).

[5] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm (2001)

[6] P. R. Chapman and G. Underwood, Visual Search of Dynamic Scenes: Event Types and the Role of Experience in Viewing Driving Situations. In: *Eye Guidance in Reading and Scene Perception*. G. Underwood (Ed.). Elsevier Science Ltd. (1998).

[7] A. T. Duchowski and A. Cöltekin,Foveated Gaze-Contingent Displays for Peripheral LOD Management, 3D Visualization, and Stereo Imaging. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 3(4):1-21 (2007).

[8] 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(11):1254–1259 (1998).

[9] W. Kienzle, B. Schölkopf, F. A. Wichmann, and M. O. Franz. How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In *Proceedings of the 29th Annual Symposium of the German Association for Pattern Recognition (DAGM 2007)*, 405–414. Springer Verlag (2007).

[10] H. L. Kundel and P. S. La Follette, Visual search patterns and experience with radiological images. *Radiology*, 103(3):523-8 (1972)

[11] G. W. McConkie and K. Rayner, The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17:578-586 (1975).

[12] 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, vol. 4662.* B. E. Rogowitz and T. N. Pappas (Eds.), 57-69 (2002).